AN INVESTIGATION OF THE PHENOTYPIC AND GENOTYPIC DETERMINANTS OF DISEASE SUSCEPTIBILITY AND PROGRESSION IN CROHN'S DISEASE

Anne Mairéad Phillips

Doctor of Philosophy
University of Edinburgh
2011

For my parents

Declaration of Originality

I declare that all work in this thesis is entirely the result of my own investigations and that, as part of a larger research group/collaboration, I have properly acknowledged the contributions of others in the section below as well as in the appropriate sections of the thesis.

The Dundee IBD recruitment (section 2.2.2) was completed with the assistance of Mrs Shirley Cleary, RN. The Dundee database was developed by Mrs Hazel Drummond and then adapted by Mrs Maureen Edwards under instruction from myself (section 2.2.6). All Edinburgh patients and controls for the database (sections 2.2.3, 2.2.4 and 2.2.5) were recruited by Mrs Linda Smith and phenotyped by Mrs Hazel Drummond with the help of other members of the Gastrointestinal Unit.

Genomic DNA extraction (section 2.3.1) was completed by the Medical Research Council Human Genetics Unit, Edinburgh. Taqman genotyping (section 2.3.3) and Illumina genotyping (section 2.3.5) was done by the Wellcome Trust Clinical Research Facility, Edinburgh and the Illumina manual reclustering was completed by Ms Angie Fawkes at the facility. Sequenom genotyping (section 1.3.4) was done at the Genomics Core of the University of California, San Francisco and financed by Genentech, Inc. All gene sequencing post initial PCR (section 2.3.10) was completed at the Medical Research Council Human Genetics Unit, Edinburgh. Fixing, mounting and sectioning of biopsy specimens (section 2.7.3) were completed by the Pathology Department of the Western General Hospital, Edinburgh and the Breakthrough Research Unit of the University of Edinburgh.

In section 6.5, the qPCR primer design was by Ms Kimberley Soo, who also completed the initial *NOD2* qPCR optimisation. Initial *GAPDH* qPCR optimisation was by Dr Marian Aldhous (section 6.5). Dr Elaine Nimmo completed the yeast-two hybrid study (section 5.2), designed *NOD2* PCR primers used in chapter 6 and selected *GALNT2* tagging SNPs (section 2.3.1). Dr Craig Stevens completed the HCT116 Western Blot in section 6.2.2.

NOD2 peptide sequence selection (section 6.2.3.1) was completed by Dr Dinesh Soares, Medical Genetics Section, Molecular Medicine Centre, and the attempt at

NOD2 antibody generation was done by Alta Bioscience, University of Birmingham (section 1.2.3.2).

Dr Nicholas Lewin-Koh, Genentech, Inc. completed the Cox Proportional Hazards modelling in sections 3.5.7 and 3.6.8 and provided statistical support for the other analyses in Chapters 3 and 4. The severity score in section 4.2.1 was developed with the consensus of Prof Jack Satsangi, Dr Ian Arnott, Dr Gwo-Tzer Ho, Dr Craig Mowat and myself.

Contents

List of Figures

List of Tables

Acknowledgements

This thesis, although my own work, would not have been possible without the support of an enormous number of people. I hope I have managed to acknowledge them all. I apologise to anyone I have unwittingly left out.

My PhD supervisors Professor Jack Satsangi and Dr Elaine Nimmo have inspired, guided and supported me through the ups and downs of the last four years. I also give grateful thanks to Dr Marian Aldhous, my unofficial third supervisor, for initially supervising experiments in the laboratory, giving me sound advice at appropriate moments (including a metaphorical kick up the backside when necessary) and generally being a great friend. Professor Cathy Abbott and Professor David Porteous, members of my thesis committee, were good sources of support and gave excellent independent advice when necessary.

I have had the privilege of working with a fantastic team of consultants, registrars, junior doctors, nurses, secretaries and support staff in the Gastrointestinal Unit at the Western General Hospital, Edinburgh. I am especially grateful to Dr Ian Arnott for useful discussions about my data, and the Endoscopy staff for their patience in letting me recruit patients for the intestinal biopsy study. Thank you to the Pathology Department of the Western General Hospital, especially Dr Paul Fineron, for help with immunohistochemistry.

I also thank everyone in the Gastrointestinal Research group at the University of Edinburgh, past and present, including Dr David Wilson, Miss Amanda Smith, Miss Rhona Aird, Dr Johan Van Limbergen, Dr Craig Stevens, Dr Paul Henderson, Dr Gwo-Tzer Ho, Dr Charlie Lees and Miss Kimberley Soo, all of whom gave me help and advice with experiments and other work. Special mention must go to Mrs Hazel Drummond who taught me how to phenotype patients and Ms Colette McColl for excellent administrative support.

I am grateful to all the technical support workers at the university who extracted and sequenced DNA (Human Genetics Unit, Medical Research Council, Edinburgh), completed the Taqman and Illumina genotyping (Wellcome Trust Clinical Research Facility, Western General Hospital, Edinburgh) and fixed/sectioned biopsy samples

(Breakthrough Breast Cancer Unit, University of Edinburgh). Dr Shona Kerr and the Generation Scotland team kindly allowed us access to the Generation Scotland Blood Donor samples.

Genentech, Inc. provided the funding for my research post; Dr Hilary Clark was the main driver there in establishing the collaboration with our group. For a most informative time at Genentech in 2009 and 2010, I thank Dr Nicholas Lewin-Koh, Dr Tushar Bhangale and Dr Hilary Clark and her research group, all of whom provided advice and help on data analysis. Thank you to the Genentech Bioinformatics Departmental head, Dr Robert Gentleman, for allowing it all to happen in the first place, and Ms Jennifer Kesler for administrative support. Working at Genentech has given me an invaluable perspective on the complex world of drug development.

In Dundee, I am particularly appreciative of the hard work of Ms Anne Andrews and Mrs Liz Chimiak, two amazing secretarial support workers, who uncomplainingly did the huge job of pulling notes for me for phenotyping. Thank you to the consultants and registrars in Dundee, especially Dr Craig Mowat and Dr Nigel Reynolds, for allowing and helping me to recruit their patients. Also thanks to Mrs Shirley Cleary for her fastidious hard work in writing to patients and helping with recruitment.

I thank my family, especially my parents, for supporting me throughout and encouraging me to believe in myself.

Professor Derek Jewell initially inspired my interest in inflammatory bowel disease while I was a clinical medical student at Oxford University. It was there that I had my initial exposure to patients with Crohn's disease and ulcerative colitis and was struck not only by how young the patients were but also how debilitating their disease could be.

Finally, an enormous thank you to all the patients, who are a constant reminder of why this work needs to be done, and have provided DNA, intestinal biopsy samples and phenotype information. I hope they see the benefit of some of my research in the future.

Abstract

The inflammatory bowel diseases (IBD), encompassing Crohn's disease (CD) and ulcerative colitis (UC), are chronic inflammatory disorders of the gastrointestinal tract. Their aetiology is not fully understood but is thought to be a combination of the effect of environmental factors in a genetically susceptible person. The work presented is an examination of the phenotypic characteristics of CD in the Scottish population, and an investigation into genetic factors that may influence susceptibility and progression.

An IBD cohort from Dundee was recruited (CD=367, UC=265), and extensive phenotypic information collected from these patients together with genomic DNA. Together with the Edinburgh CD cohort already established, the total CD population (n=1155) was examined for time to disease progression (stricturing and/or penetrating disease, according to the Montreal classification) and first resection; a multivariate analysis was performed for factors influencing these outcomes. In this Scottish CD population, the median time to disease progression and first resection was 14.2 years and 8.9 years respectively. The major factor influencing risk of resection and disease progression was disease location, with patients having pure ileal (L1) disease or ileocolonic (L3) disease being more susceptible than those with pure colonic (L2) disease. Compared with L2 disease, the hazards ratios (HR) for disease progression were 4.7 and 2.8, and risk of resection 5.2 and 2.6 for L1 and L3 disease respectively.

Disease progression and risk of resection are surrogate markers of disease severity. To try to better understand the determinants of severe disease, a novel score for disease severity was developed and applied to the Dundee CD cohort. This composite score encompassed the variables of medical and surgical management, disease behaviour and location, nutritional status as well as hospitalisations, with a total score that could range from 1 to 16. A score of 7 or more was found to define the 50% of patients with the most severe disease. This cut-off was used to divide patients into less severe and more severe categories; phenotypic and genetic factors were examined for correlation with more severe disease. Genetic factors examined were the 32 most significant CD susceptibility single nucleotide polymorphisms

(SNPs) uncovered by recent genome-wide association scans (GWAS). Factors correlated with more severe disease included disease location (L1, odds ratio (OR) 2.20, p=0.0025), age group at diagnosis (p=0.0004) and two CD susceptibility SNPs (rs9286879 and rs17582416; p=0.0085 and p=0.045 respectively).

NOD2 was the first IBD susceptibility gene identified. In order to further define pathways involving *NOD2*, a yeast two-hybrid screen in our laboratory using *NOD2* cDNA as the bait had already identified an interaction between *NOD2* and UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase (*GALNT2*). This enzyme is involved in O-glycosylation, important in the post translational modification of mucins. A *GALNT2* genotype/phenotype analysis on the Edinburgh IBD population was completed, with the Dundee IBD population used as a replication cohort. In the Edinburgh IBD population, the *GALNT2* tagging SNP rs7536663 was associated with CD susceptibility (OR 1.38, p=0.0008 vs controls), but replication was not achieved in the Dundee cohort (p=0.469). There was no association of any of the *GALNT2* SNPs with UC.

The *GALNT2/NOD2* interaction was further investigated by completing co-immunoprecipitation between the two genes to characterise the level and type of interaction. An interaction between *GALNT2* and *NOD2* was confirmed in mammalian cells, with the interaction being at the N-terminal end of the *NOD2* protein. *GALNT2* expression in a cell line and biopsies was investigated by quantitative polymerase chain reaction and immunohistochemistry respectively. There were no statistically significant changes in *GALNT2* or *NOD2* mRNA expression in the LS174T cell line after stimulation with specific ligands for *NOD2* and *GALNT2*. GALNT2 protein expression was characterised in intestinal biopsy samples to be predominantly in the lamina propria, with some expression in the enterocytes.

To further define the contribution of mucin genes to IBD susceptibility, tagging SNPs across the *MUC2*, *MUC3A* and *MUC19* genes were genotyped in the Edinburgh IBD cohort and examined for a link with IBD, CD and UC susceptibility, but associations were not found. In view of the strong association with CD susceptibility of a SNP near the *MUC19* locus in a recent GWAS, tagging SNPs across the leucine rich repeat kinase-2 (*LRRK2*) gene, near the *MUC19* gene, were also genotyped and examined in the Dundee cohort for an association with IBD, CD and UC susceptibility, but was also negative when corrected for multiple testing.

The studies presented allow an improved understanding of the influence of phenotypic characteristics on disease progression, need for surgery and severity in CD. The role of disease location has been determined to be particularly critical, in keeping with other published studies. A detailed examination of the influence of specific genes on disease susceptibility has failed to definitely demonstrate an association between germline variation in *GALNT2*, *MUC2*, *MUC3A*, *MUC19* or *LRRK2* and IBD, CD or UC susceptibility. An interaction in mammalian cells between NOD2 and GALNT2 has been shown, but further work is required to demonstrate that this is a biologically relevant interaction.

List of Publications arising directly from this thesis

Pharmacogenetics and Inflammatory Bowel Disease
Phillips, A.M., Hare, N., Satsangi, J.
Current Pharmacogenomics and Personalized Medicine 2008; 6(3):201-224

Detailed haplotype-tagging study of germline variation of MUC19 in Inflammatory Bowel Disease
Phillips, A.M., Nimmo, E.R., Van Limbergen, J., Drummond, H., Smith, L., and Satsangi, J.
Inflammatory Bowel Diseases (2009); 16(4):557-558

Analysis of protein-protein and gene-gene interactions implicates TLE1 as a critical modifier of NOD2 effect in Crohn's disease
Nimmo, E.R., Phillips, A.M., Stevens, C., Smith, A., Drummond, H.E., Noble, C.L., Quail, M., Davies, G., Aldhous, M.C., Wilson, D.C., Satsangi, J.
Gastroenterology (2011) in press

List of Publications arising from patient recruitment occurring as part of this thesis

Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47
Anderson, C.A., Boucher, G., Lees, C.W., Franke, A., D'Amato, M., Taylor, K.D., Lee, J.C., Goyette, P., Imielinski, M., Latiano, A., Lagace, C., Scott, R., Amininejad, L., Bumpstead, S., Baidoo, L., Baldassano, R.N., Barclay, M., Bayless, T.M., Brand, S., Buning, C., Colombel, J.F., Denson, L.A., de, Vos M., Dubinsky, M., Edwards, C., Ellinghaus, D., Fehrmann, R.S., Floyd, J.A., Florin, T., Franchimont, D., Franke, L., Georges, M., Glas, J., Glazer, N.L., Guthery, S.L., Haritunians, T., Hayward, N.K., Hugot, J.P., Jobin, G., Laukens, D., Lawrance, I., Lemann, M., Levine, A., Libioulle, C., Louis, E., McGovern, D.P., Milla, M., Montgomery, G.W., Morley, K.I., Mowat, C., Ng, A., Newman, W., Ophoff, R.A., Papi, L., Palmieri, O., Peyrin-Biroulet, L., Panes, J., **Phillips, A.**, Prescott, N.J., Proctor, D.D., Roberts, R., Russell, R., Rutgeerts, P., Sanderson, J., Sans, M., Schumm, P., Seibold, F., Sharma, Y., Simms, L.A., Seielstad, M., Steinhart, A.H., Targan, S.R., van den Berg, L.H., Vatn, M., Verspaget, H., Walters, T., Wijmenga, C., Wilson, D.C., Westra, H.J., Xavier, R.J., Zhao, Z.Z., Ponsioen, C.Y., Andersen, V., Torkvist, L., Gazouli, M., Anagnou, N.P., Karlsen, T.H., Kupcinskas, L., Sventoraityte, J., Mansfield, J.C., Kugathasan, S., Silverberg, M.S., Halfvarson, J., Rotter, J.I., Mathew, C.G., Griffiths, A.M., Gearry, R., Ahmad, T., Brant, S.R., Chamaillard, M., Satsangi, J., Cho, J.H., Schreiber, S., Daly, M.J., Barrett, J.C., Parkes, M., Annese, V., Hakonarson, H., Radford-Smith, G., Duerr, R.H., Vermeire, S., Weersma, R.K., Rioux, J.D.
Nature Genetics (2011); 43(3): 246-252

Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci
Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., Anderson, C. A., Bis, J. C., Bumpstead, S., Ellinghaus, D., Festen, E. M., Georges, M., Green, T., Haritunians, T., Jostins, L., Latiano, A., Mathew, C. G., Montgomery, G. W., Prescott, N. J., Raychaudhuri, S., Rotter, J. I., Schumm, P., Sharma, Y., Simms, L. A., Taylor, K. D., Whiteman, D., Wijmenga, C., Baldassano, R. N., Barclay, M., Bayless, T. M., Brand, S., Buning, C., Cohen, A., Colombel, J. F., Cottone, M., Stronati, L., Denson, T., de, Vos M., D'Inca, R., Dubinsky, M., Edwards, C., Florin, T., Franchimont, D., Gearry, R., Glas, J., Van, Gossum A., Guthery, S. L., Halfvarson, J., Verspaget, H. W., Hugot, J. P., Karban, A., Laukens, D., Lawrance, I., Lemann, M., Levine, A., Libioulle, C., Louis, E., Mowat, C., Newman, W., Panes, J., **Phillips, A**., Proctor, D. D., Regueiro, M., Russell, R., Rutgeerts, P., Sanderson, J., Sans, M., Seibold, F., Steinhart, A. H., Stokkers, P. C., Torkvist, L., Kullak-Ublick, G., Wilson, D., Walters, T., Targan, S. R., Brant, S. R., Rioux, J. D., D'Amato, M., Weersma, R. K., Kugathasan, S., Griffiths, A. M., Mansfield, J. C.,

Vermeire, S., Duerr, R. H., Silverberg, M. S., Satsangi, J., Schreiber, S., Cho, J. H., Annese, V., Hakonarson, H., Daly, M. J., and Parkes, M.
Nature Genetics (2010); 42(12):1118-1125

Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls
Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D. F., Giannoulatou, E., Holmes, C., Marchini, J. L., Stirrups, K., Tobin, M. D., Wain, L. V., Yau, C., Aerts, J., Ahmad, T., Andrews, T. D., Arbury, H., Attwood, A., Auton, A., Ball, S. G., Balmforth, A. J., Barrett, J. C., Barroso, I., Barton, A., Bennett, A. J., Bhaskar, S., Blaszczyk, K., Bowes, J., Brand, O. J., Braund, P. S., Bredin, F., Breen, G., Brown, M. J., Bruce, I. N., Bull, J., Burren, O. S., Burton, J., Byrnes, J., Caesar, S., Clee, C. M., Coffey, A. J., Connell, J. M., Cooper, J. D., Dominiczak, A. F., Downes, K., Drummond, H. E., Dudakia, D., Dunham, A., Ebbs, B., Eccles, D., Edkins, S., Edwards, C., Elliot, A., Emery, P., Evans, D. M., Evans, G., Eyre, S., Farmer, A., Ferrier, I. N., Feuk, L., Fitzgerald, T., Flynn, E., Forbes, A., Forty, L., Franklyn, J. A., Freathy, R. M., Gibbs, P., Gilbert, P., Gokumen, O., Gordon-Smith, K., Gray, E., Green, E., Groves, C. J., Grozeva, D., Gwilliam, R., Hall, A., Hammond, N., Hardy, M., Harrison, P., Hassanali, N., Hebaishi, H., Hines, S., Hinks, A., Hitman, G. A., Hocking, L., Howard, E., Howard, P., Howson, J. M., Hughes, D., Hunt, S., Isaacs, J. D., Jain, M., Jewell, D. P., Johnson, T., Jolley, J. D., Jones, I. R., Jones, L. A., Kirov, G., Langford, C. F., Lango-Allen, H., Lathrop, G. M., Lee, J., Lee, K. L., Lees, C., Lewis, K., Lindgren, C. M., Maisuria-Armer, M., Maller, J., Mansfield, J., Martin, P., Massey, D. C., McArdle, W. L., McGuffin, P., McLay, K. E., Mentzer, A., Mimmack, M. L., Morgan, A. E., Morris, A. P., Mowat, C., Myers, S., Newman, W., Nimmo, E. R., O'Donovan, M. C., Onipinla, A., Onyiah, I., Ovington, N. R., Owen, M. J., Palin, K., Parnell, K., Pernet, D., Perry, J. R., **Phillips, A**., Pinto, D., Prescott, N. J., Prokopenko, I., Quail, M. A., Rafelt, S., Rayner, N. W., Redon, R., Reid, D. M., Renwick, Ring, S. M., Robertson, N., Russell, E., St, Clair D., Sambrook, J. G., Sanderson, J. D., Schuilenburg, H., Scott, C. E., Scott, R., Seal, S., Shaw-Hawkins, S., Shields, B. M., Simmonds, M. J., Smyth, D. J., Somaskantharajah, E., Spanova, K., Steer, S., Stephens, J., Stevens, H. E., Stone, M. A., Su, Z., Symmons, D. P., Thompson, J. R., Thomson, W., Travers, M. E., Turnbull, C., Valsesia, A., Walker, M., Walker, N. M., Wallace, C., Warren-Perry, M., Watkins, N. A., Webster, J., Weedon, M. N., Wilson, A. G., Woodburn, M., Wordsworth, B. P., Young, A. H., Zeggini, E., Carter, N. P., Frayling, T. M., Lee, C., McVean, G., Munroe, P. B., Palotie, A., Sawcer, S. J., Scherer, S. W., Strachan, D. P., Tyler-Smith, C., Brown, M. A., Burton, P. R., Caulfield, M. J., Compston, A., Farrall, M., Gough, S. C., Hall, A. S., Hattersley, A. T., Hill, A. V., Mathew, C. G., Pembrey, M., Satsangi, J., Stratton, M. R., Worthington, J., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W., Parkes, M., Rahman, N., Todd, J. A., Samani, N. J., and Donnelly, P.
Nature (2010); 464(7289):713-720

Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls
Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D. F., Giannoulatou, E., Holmes, C., Marchini, J. L., Stirrups, K., Tobin, M. D., Wain, L. V., Yau, C., Aerts, J., Ahmad, T., Andrews, T. D., Arbury, H., Attwood, A., Auton, A., Ball, S. G., Balmforth, A. J., Barrett, J. C., Barroso, I., Barton, A., Bennett, A. J., Bhaskar, S., Blaszczyk, K., Bowes, J., Brand, O. J., Braund, P. S., Bredin, F., Breen, G., Brown, M. J., Bruce, I. N., Bull, J., Burren, O. S., Burton, J., Byrnes, J., Caesar, S., Clee, C. M., Coffey, A. J., Connell, J. M., Cooper, J. D., Dominiczak, A. F., Downes, K., Drummond, H. E., Dudakia, D., Dunham, A., Ebbs, B., Eccles, D., Edkins, S., Edwards, C., Elliot, A., Emery, P., Evans, D. M., Evans, G., Eyre, S., Farmer, A., Ferrier, I. N., Feuk, L., Fitzgerald, T., Flynn, E., Forbes, A., Forty, L., Franklyn, J. A., Freathy, R. M., Gibbs, P., Gilbert, P., Gokumen, O., Gordon-Smith, K., Gray, E., Green, E., Groves, C. J., Grozeva, D., Gwilliam, R., Hall, A., Hammond, N., Hardy, M., Harrison, P., Hassanali, N., Hebaishi, H., Hines, S., Hinks, A., Hitman, G. A., Hocking, L., Howard, E., Howard, P., Howson, J. M., Hughes, D., Hunt, S., Isaacs, J. D., Jain, M., Jewell, D. P., Johnson, T., Jolley, J. D., Jones, I. R., Jones, L. A., Kirov, G., Langford, C. F., Lango-Allen, H., Lathrop, G. M., Lee, J., Lee, K. L., Lees, C., Lewis, K., Lindgren, C. M., Maisuria-Armer, M., Maller, J., Mansfield, J., Martin, P., Massey, D. C., McArdle, W. L., McGuffin, P., McLay, K. E., Mentzer, A., Mimmack, M. L., Morgan, A. E., Morris, A. P., Mowat, C., Myers, S., Newman, W., Nimmo, E. R., O'Donovan, M. C., Onipinla, A., Onyiah, I., Ovington, N. R., Owen, M. J., Palin, K., Parnell, K., Pernet, D., Perry, J. R., **Phillips, A.**, Pinto, D., Prescott, N. J., Prokopenko, I., Quail, M. A., Rafelt, S., Rayner, N. W., Redon, R., Reid, D. M., Renwick, Ring, S. M., Robertson, N., Russell, E., St, Clair D., Sambrook, J. G., Sanderson, J. D., Schuilenburg, H., Scott, C. E., Scott, R., Seal, S., Shaw-Hawkins, S., Shields, B. M., Simmonds, M. J., Smyth, D. J., Somaskantharajah, E., Spanova, K., Steer, S., Stephens, J., Stevens, H. E., Stone, M. A., Su, Z., Symmons, D. P., Thompson, J. R., Thomson, W., Travers, M. E., Turnbull, C., Valsesia, A., Walker, M., Walker, N. M., Wallace, C., Warren-Perry, M., Watkins, N. A., Webster, J., Weedon, M. N., Wilson, A. G., Woodburn, M., Wordsworth, B. P., Young, A. H., Zeggini, E., Carter, N. P., Frayling, T. M., Lee, C., McVean, G., Munroe, P. B., Palotie, A., Sawcer, S. J., Scherer, S. W., Strachan, D. P., Tyler-Smith, C., Brown, M. A., Burton, P. R., Caulfield, M. J., Compston, A., Farrall, M., Gough, S. C., Hall, A. S., Hattersley, A. T., Hill, A. V., Mathew, C. G., Pembrey, M., Satsangi, J., Stratton, M. R., Worthington, J., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W., Parkes, M., Rahman, N., Todd, J. A., Samani, N. J., and Donnelly, P.
Nature (2010); 464(7289):713-720

Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region
Barrett, J. C., Lee, J. C., Lees, C. W., Prescott, N. J., Anderson, C. A., **Phillips, A.**, Wesley, E., Parnell, K., Zhang, H., Drummond, H., Nimmo, E. R., Massey, D., Blaszczyk, K., Elliott, T., Cotterill, L., Dallal, H., Lobo, A. J., Mowat, C., Sanderson, J. D., Jewell, D. P., Newman, W. G., Edwards, C., Ahmad, T., Mansfield, J. C.,

Satsangi, J., Parkes, M., Mathew, C. G., Donnelly, P., Peltonen, L., Blackwell, J. M., Bramon, E., Brown, M. A., Casas, J. P., Corvin, A., Craddock, N., Deloukas, P., Duncanson, A., Jankowski, J., Markus, H. S., Mathew, C. G., McCarthy, M. I., Palmer, C. N., Plomin, R., Rautanen, A., Sawcer, S. J., Samani, N., Trembath, R. C., Viswanathan, A. C., Wood, N., Spencer, C. C., Barrett, J. C., Bellenguez, C., Davison, D., Freeman, C., Strange, A., Donnelly, P., Langford, C., Hunt, S. E., Edkins, S., Gwilliam, R., Blackburn, H., Bumpstead, S. J., Dronov, S., Gillman, M., Gray, E., Hammond, N., Jayakumar, A., McCann, O. T., Liddle, J., Perez, M. L., Potter, S. C., Ravindrarajah, R., Ricketts, M., Waller, M., Weston, P., Widaa, S., Whittaker, P., Deloukas, P., Peltonen, L., Mathew, C. G., Blackwell, J. M., Brown, M. A., Corvin, A., McCarthy, M. I., Spencer, C. C., Attwood, A. P., Stephens, J., Sambrook, J., Ouwehand, W. H., McArdle, W. L., Ring, S. M., and Strachan, D. P.
Nature Genetics (2009); 41(12):1330-1334

List of Published Abstracts

Clinical and genetic factors predict severe disease: a novel composite index
Phillips, A.M., Arnott, I., Heron, T., Mowat, C., Clark, H., Lewin-Koh, N.
Poster presentation, British Society of Gastroenterology Annual Meeting, March 2011
Poster Presentation, Digestive Diseases Week, May 2011
Poster Presentation, British Society of Gastroenterology Annual Meeting, March 2011

Surgery in Crohn's disease: Analysis of 15000 patient-years experience in Eastern Scotland
Phillips, A.M., Cleary, S., Smith, L., Drummond, H.E., Lewin-Koh, N., Clark, H., Mowat, C., Arnott, I.D., Satsangi, J.
Poster Presentation, Digestive Diseases Week, May 2011

Risk of complications in a Scottish Crohn's disease cohort and association with disease location
Phillips, A.M., Aldhous, M.C., Lewin-Koh, N., McLeod, S., Smith, L., Drummond, H.E., Pidasheva, S., Mowat, C., Clark, H. and Satsangi, J.
Oral Presentation, British Society of Gastroenterology Annual Meeting, March 2010

Germline variation of a novel NOD2/CARD15 interacting protein, GALNT2, is associated with genetic susceptibility to Crohn's disease
Phillips, A.M., Van Limbergen, J., Davies, G., Drummond, H.E., Smith, L.A., Smith, A.J., Satsangi, J. and Nimmo, E.R.
Poster presentation, Digestive Diseases Week, May 2009

Detailed haplotype-tagging study of germline variation of MUC19 in Inflammatory Bowel Disease
Phillips, A.M., Nimmo, E.R., Van Limbergen, J., Drummond, H.E., Smith, L.A., Satsangi, J.
Poster presentation, Digestive Diseases Week, May 2009
Poster presentation, British Society of Gastroenterology Annual Meeting, March 2009

Abbreviations

5ASA	5-aminosalicylate
A	Adenine
ABCB	ATP-binding cassette, subfamily B
AIEC	Adherent-invasive Escherichia coli
ASO	Allele specific oligonucleotide
ATG	Autophagy
ATG16L1	Autophagy 16-Like 1
BMI	Body Mass Index
C	Cytosine
C11ORF30	Chromosome open reading frame 30
CARD	Caspase recruitment domain-containing protein
CD	Crohn’s disease
CDKAL1	CDK regulatory subunit-associated protein1-like1
cDNA	Complementary DNA
Chr	Chromosome
CI	Confidence interval
CMV	Cytomegalovirus
CNV	Copy number variation
CoIP	Co-immunoprecipitation
DLG5	Homolog of Discs large drosophilia, 5
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dNTP	Deoxyribonucleotide triphosphate
DSS	Dextran sodium sulfate
Dx	Diagnosis
DZ	Dizygotic
EDTA	Ethylenediaminetetraacetic acid
FCS	Fetal calf serum
Freq	Frequency
G	Guanine
GalNAc	N-acetyl-D-galactosamine
GALNT2	UDP-N-acetyl-α-D-galactosamine polypeptide N-acetylgalactosaminyltransferase
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GI	Gastrointestinal
GWAS	Genome-wide association study
HA	Haemagglutinin
HCl	Hydrochloric acid
HIER	Heat Induced Epitope Retrieval
HLA	Human Leukocyte Antigen

HR	Hazard ratio
HRP	Horse radish peroxidase
HW	Hardy Weinberg
IBD	Inflammatory bowel disease
ICOSLG	Inducible T-cell costimulator ligand
IFN	Interferon
IHC	Immunohistochemistry
IL	Interleukin
IL23R	Interleukin-23 receptor
IPTG	Isopropyl-1-thio-β-D-galactopyranoside
IRGM	Immunity-related p47 guanosine triphosphatase family, M
ITLN1	Intelectin 1
JAK	Janus kinase
kDa	Kilo daltons
KLH	Keyhole limpet haemocynin
KPNA7	Karyopherin alpha 7
LB	Lysogeny broth
LC3	Light chain 3
LD	Linkage disequilibrium
LPS	Lipopolysaccharide
LREC	Local research ethics committee
LRR	Leucine rich repeat
LRRK2	Leucine rich repeat kinase-2
LSO	Locus specific oligonucleotide
MAF	Minor allele frequency
MALDI-TOF	Matrix-assisted laser desorption/ionization-time of flight
MAP	Mycobacterium avium subspecies paratuberculosis
MDP	Muramyl dipeptide
MDR1	Multidrug resistance 1
MEM	Minimum essential medium eagle
$MgCl_2$	Magnesium chloride
MHC	Major histocompatibility complex
MRC	Medical Research Council
MST1	Macrophage stimulating 1
MUC	Mucin
Myc	Myelcytomatosis
MZ	Monozygotic
NaCl	Sodium chloride
NBD	Nucleotide binding domain
NF-κB	Nuclear factor kappa-light chain enhancer of activated B cells
NKX2-3	NK2 Homeobox 3
NOD	Nucleotide-binding oligomerization domain protein

NPV	Negative predictive value
OCTN	Organic cation transporter
OMAP	Octomeric multiple antigenic peptide
OmpC	Outer membrane porin C
OR	Odds Ratio
ORMDL3	ORM-like protein 3
PARK8	Parkinson disease 8
PBS	Phosphate buffered saline
pCMV	Plasmid cytomegalovirus
PCR	Polymerase chain reaction
PD	Parkinson's disease
PPAR- γ	peroxisome proliferator-activated receptor gamma
PPV	Positive predictive value
PTGER4	Prostglandin E receptor 4
PTPN	Protein-tyrosine phosphatise, nonreceptor type
PVDF	Polyvinylidene Fluoride
qPCR	Quantitative polymerase chain reaction
R^2	Coefficient of determination
RFLP	Restriction fragment length polymorphism
RIP2/RICK	Receptor interacting protein 2, RIP-like interacting Clarp kinase
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
Rpm	Revolutions per minute
rRNA	Ribosomal ribonucleic acid
RT-PCR	Real time polymerase chain reaction
SCFA	Short-chain fatty acids
SEM	Standard error of the mean
Ser	Serine
SLC	Solute carrier family
SMURF1	SMAD-specific E3 ubiquiting protein ligase 1
SNP	Single nucleotide polymorphism
STAT	Signal transducer and activator of transcription
T	Thymine
Taq	Taq polymerase
TBE	Tris/Borate/EDTA
Thr	Threonine
TLR	Toll-like receptor
TNFα	Tumour necrosis factor alpha
TNFSF15	Tumour necrosis factor ligand superfamily, member 15
UC	Ulcerative colitis
UTR	Untranslated region
v/v	Volume per volume

VNTR	Variable number of tandem repeats
w/v	Weight per volume
wt	Wild type
WTCCC	Wellcome Trust Case Control Consortium
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
ZNF365	Zinc finger protein 365

Chapter 1 Introduction

1.1 The inflammatory bowel diseases

Crohn's disease (CD; MIM 266600) and ulcerative colitis (UC; MIM 191390) are chronic inflammatory disorders of the gastrointestinal tract of incompletely understood aetiology and for which there are no known cures.

1.1.1 Pathology and presentation

CD is characterised by patchy transmural inflammation that can affect any part of the gastrointestinal tract from the mouth to the anus, but most commonly affects the ileocaecal region. There is a tendency to develop inflammatory or fibrotic strictures as well as fistulae, both internal and perianal. The presentation of the disease can be rather heterogeneous due to the variety of locations that can be affected. Typically, symptoms can include weight loss, abdominal pain, diarrhoea, vomiting and systemic upset.

UC is characterised by a transmucosal inflammation that is continuous from the rectum to the extent of the disease, but classically only affects the colon, although the terminal ileum can be affected by a backwash ileitis, and in very rare cases a more diffuse small bowel inflammation can result. The predominant symptom tends to be bloody diarrhoea.

There is a multitude of extraintestinal manifestations affecting both CD and UC patients, including uveitis, erythema nodosum, primary sclerosing cholangitis, and in patients with colonic disease, a long-term increased risk of colon cancers.[1]

1.1.2 Diagnosis

The diagnosis of inflammatory bowel disease (IBD) is by the Lennard-Jones criteria[2], which states exclusion and inclusion criteria for both CD and UC. UC inclusion is continuous mucosal inflammation without granulomata affecting the rectum and some or the entire colon in continuity with the rectum, with specific exclusion criteria including infective colitis, ischaemic colitis and abnormalities suggesting CD (e.g. small bowel disease). CD inclusion is discontinuous inflammation in areas from the mouth to the anus, with transmural inflammation with granulomata, having the potential for fibrosis, abscesses, fistulae and fissuring ulcers. Exclusion criteria for CD include infection and ischaemia.

1.1.3 Incidence and Prevalence

The incidence of IBD has been widely studied, with age-adjusted incidence rates varying from 5-20 people/100,000/year for CD, and 3-20 people/100,000/year for UC, as shown in Table 1-1. Point prevalence of CD is around 150/100,000 and UC 200/100,000, as shown in Table 1-1. There has been debate as to whether the incidence and prevalence is rising, and whether these rates are higher in countries with more northerly latitude. A study in Europe[3] looking at new diagnoses with IBD between 1991 and 1993 at 20 European centres found that there was a higher incidence in northern compared with southern centres. Through the middle of the 20th century the incidence of IBD was rising, but it is now thought to have plateaued.[4;5] Whether the rise was due to better diagnosis of the diseases or a rise in incidence is not clear.

Country	Region	Reference	Years	CD incidence /100,000/yr	UC incidence /100,000/yr	CD prevalence /100,000	UC prevalence /100,000
Canada		Bernstein[6]	1998-2000	8.8-20.2	9.9-19.5	161-319	162-249
USA	Olmsted, Minnesota	Loftus[5]	1990-2000	7.9	8.8	174	214
UK	Derby	Fellows[4]	1976-1985	6.67	NS	85	NS
Sweden	Örebro	Lindberg[7]	1963-1987	6.1	NS	146	NS
Scotland	Tayside	Steed[8]	2003-2007	9.56	NS	157	NS
France	Nord-Pas de Calais and Somme	Gower-Rousseau[9]	1988-1990	4.9	3.2	NS	NS

Table 1-1 IBD incidence and prevalence, NS= not studied

1.1.3.1 Age at incidence

The peak age of incidence of IBD is universally recognised to be between 20 and 30 years, when the age-specific incidence of the disease is 10-40/100,000[6;9-12], with a possible second peak between 60-70 years.[5;13] Thus IBD is predominantly a disease that begins in early adulthood.

1.1.4 Classification

1.1.4.1 Crohn's disease

Due to its heterogeneous nature, it is important to subclassify CD into disease types for research and clinical purposes. Initial attempts to classify CD were on the basis of location of disease[14], but it was later recognised that disease behaviour was also an important variable, as patients with penetrating disease appeared to have a different disease course to those who did not.[15] The resulting 'Rome classification' had disease location variables (stomach/duodenum, jejunum, ileum, colon, rectum, anal/perianal) along with behaviour variables (primarily inflammatory, fistulating or fibrostenotic), disease extent (localised or diffuse) and operative history (primary or recurrent)[16], but was not widely adopted because of its complexity. The Vienna classification[17] simplified this classification, using 3 variables: age at diagnosis (below 40 years old: A1, or 40 and above: A2), location (terminal ileum: L1, colon: L2, ileocolon: L3 and upper GI: L4) and behaviour (non stricturing non penetrating: B1, stricturing: B2, penetrating: B3). The location variable was the maximum extent before the first resection, and the 4 locations were mutually exclusive; thus patients with upper GI disease and disease elsewhere were classified as being in the L4 category. As only a small proportion of L4 patients have disease localised *only* to the upper GI tract, valuable phenotypic information was lost. In addition, the age classification cut-off of 40 years old at diagnosis did not take full account of the peak age of onset of disease, meaning that the majority of patients were in the A1 age category. In the behaviour category, no time limit was put on the behaviour variable. In addition, perianal disease was counted as penetrating disease; evidence subsequently emerged that perianal disease was phenotypically quite different from internally penetrating disease.[18;19]

These shortcomings were addressed in the Montreal classification[20], which reclassified B3 disease as referring only to internally penetrating disease, but allowed any of the B1-B3 disease behaviour variables to be suffixed by 'p' to indicate that perianal disease was present. It suggested that the behaviour variable should be assessed at the 5 year time point. The classification also allowed L4 to be classified along with other locations, and replaced the two age categories with three: less than 17 years old at diagnosis, 17-40 at diagnosis and more than 40 at diagnosis.

Recently, a modification of the Montreal classification ('Paris classification') has been agreed for paediatric patients.[21] The A1 age group for paediatric onset patients has been further subdivided into A1a (age of onset less than 10 years old) and A1b (age of onset between 10 and 17 years old). In addition, the L4 location has been subdivided into proximal upper GI, L4a (oesophageal, gastric or duodenal disease) and distal upper GI, L4b (jejunal, proximal and mid ileal disease). Finally, the Paris classification has also added a B2B3 category to note when stricturing and penetrating disease co-exist. At present these modifications apply to paediatric-onset disease only.

1.1.4.2 Ulcerative colitis

UC was first formally classified in the Montreal classification[20] by extent (proctitis E1; left sided disease E2; and disease proximal to the splenic flexure E3) and current severity (S0 asymptomatic; S1 mild; S2 moderate and S3 severe ulcerative colitis). The recent Paris classification of paediatric inflammatory bowel disease[21] has modified the E3 category to disease extending past the splenic flexure but no further than the hepatic flexure, with a new E4 category denoting disease proximal to the hepatic flexure. At present this modification applies to paediatric-onset disease only.

1.2 Management of IBD

The general aims of treatment for IBD are to alleviate symptoms and to improve or remove inflammation, strictures and fistulae. Whether early medical therapy in the form of thiopurines or biological therapy can prevent or delay long term disease progression, rather than just prevent relapse, is a subject of ongoing debate.

1.2.1 Medical management

1.2.1.1 Corticosteroids

One of the first ever double-blind placebo controlled trials in medicine proved the benefit of steroids in UC in the acute setting.[22;23] Since then, steroids, both systemic (oral and intravenous) and topical, have been the mainstay of treatment in IBD, and in the past they have been used as maintenance therapy. In more recent years there has been recognition of their long term risks and the need to use other 'steroid-sparing' therapies in the medium to long term. However, they still have a place in IBD management on a short term basis in order to induce remission in patients with active disease.

1.2.1.2 5-Aminosalicylates

The mechanism of action of 5-aminosalicylates (5-ASA) in IBD is thought to be through activation of the γ-form of the peroxisome proliferator-activated receptors (PPAR-γ) which are highly expressed in the colon and cause modulation of cytokine production.[24]

The 5ASA compounds have a well-established role in the maintenance of remission in UC[25], and may help to prevent development of colorectal cancers in this population.[26] 5ASA compounds are also important in the induction of remission in UC, both systemically[27] and topically.[28] A recent Cochrane meta-analysis found that there was no benefit of 5ASA compounds over placebo in the maintenance of remission of CD[29], although they are still frequently prescribed to patients with colonic disease.

1.2.1.3 Thiopurines

The prodrug azathioprine and its active metabolite 6-mercaptopurine are purine antagonists that compete in biochemical processes requiring purines (e.g. DNA and RNA synthesis). The exact mechanisms whereby thiopurines exert their immunosuppressive effect in IBD are poorly understood, but include promotion of the apoptosis of activated T lymphocytes.[30] Thiopurines have a well established role in the induction of remission of CD, with an odds ratio (OR) of 2.43 compared with placebo for induction of remission.[31] They are also known to be beneficial in the maintenance of remission, with an OR of 2.32 compared with placebo. Their role in UC is less clear. Although they are of benefit in the maintenance of remission, with an OR of 0.41 vs. placebo for failure to maintain remission[32], there are conflicting data as to the benefit of thiopurines in the induction of remission, with a meta-analysis not conclusively proving their benefit.[33]

1.2.1.4 Anti-TNFα therapies

Tumour necrosis factor alpha (TNFα) is a key cytokine involved in stimulating the acute phase reaction, and is involved in the pathogenesis of many inflammatory diseases. Serum TNFα levels have been shown to be elevated in IBD patients.[34] Infliximab is a chimeric human-murine monoclonal antibody which inhibits TNFα activity. Initial studies focused on its use in CD patients. Following a series of small studies showing benefit[35-37], a larger trial involving over 500 patients (ACCENT–I) demonstrated that patients with poorly controlled CD were more likely to achieve clinical remission than patients who had placebo, and that it was effective at maintaining remission.[38] More recently, studies in patients with acute severe UC have also shown benefit.[39] However, up to a third of patients fail to respond and in many of those who do, the effect can be short-lived. There is evidence to suggest that the formation of antibodies against infliximab is correlated with an increased risk of infusion reactions and reduced duration of response to treatment.[40] Theoretically adalimumab, as a fully humanized anti-TNFα antibody, should offer a reduced risk of infusion reactions and delayed hypersensitivity reactions. Compared to infliximab, very similar remission rates were seen with adalimumab in the CLASSIC-I trial in patients naïve to anti TNF therapy.[41] The rate of antibody formation was low (0.04%) but as the study was only 4 weeks long this figure is likely to be an underestimate. A follow-on study, CLASSIC-II, compared adalimumab and placebo in the maintenance of remission[42] and again produced comparable results to infliximab, with an antibody rate of 2.6%, compared to 14% with infliximab in ACCENT-I. A larger study, CHARM, supported these remission rates but did not look at immunogenicity.[43] Only one study so far has looked specifically at the use of adalimumab in patients who have had infusion reactions with infliximab, or have lost response to it.[44] Adalimumab has the added benefit over infliximab of being given subcutaneously rather than intravenously, allowing home self-administration.

Apart from the short term problems of infusion reactions and hypersensitivity reactions with biologicals, other side effects can be more serious. The risk of infection may be increased with patients who receive anti-TNFα therapies[45] and previous tuberculosis is a relative contraindication due to risk of reactivation of the disease. Case reports of patients developing haemopoetic cancers[46] and solid tumours[47-49] following anti TNFα therapies have rightly raised concerns about the long term risks of anti TNF agents, especially when used in combination with immunosuppressants including azathioprine. However, more recent evidence from the SONIC trial has indicated that in moderate-to-severe CD, combination therapy with thiopurines and infliximab is more likely to bring patients into, and maintain them in, steroid-free remission.[50] The long term safety of this strategy is unclear.

1.2.1.5 Novel therapies

There is a clinical need to develop novel therapies to complement the existing repertoire. However, such advances have been slow. A clinical trial of intravenous alicaforsen, an antisense inhibitor of intercellular adhesion molecule 1 (ICAM-1), failed to demonstrate any clinical benefit in patients with CD.[51] An anti-p40 antibody (ABT-894) against the common p40 subunit of IL12 and IL23 has shown promise in a small clinical trial in 79 patients with active CD.[52] Another anti-p40 monoclonal antibody, ustekinumab, was used in 104 CD patients and showed an early benefit from the treatment[53]; a larger trial is underway.

Autologous stem cell transplantation, which in theory could 'reset' the immune system has shown promise in a non-randomised study in 12 patients with refractory CD[54]; further randomised trials are ongoing.

1.2.2 Surgical management

1.2.2.1 Ulcerative colitis

Colectomy can be indicated for acute severe disease, low grade continuous inflammation refractory to medical therapy, or development of colorectal cancer or colonic high grade dysplasia. In one study the proportion of patients who had required colectomy was 20% at 5 years after diagnosis and 45% by 25 years; the main factor affecting need for colectomy was disease extent.[55]

In patients admitted with acute severe disease (as defined by the Truelove-Witts criteria[23]), 30-40% will require colectomy on that admission[56;57], with, on day 3, a stool frequency of >8/day, or CRP>45mg/l and stool frequency of 3-8, having a 85% risk of colectomy on that admission.[57] An alternative risk score[56] cites colonic dilatation and/or a mean stool frequency >9 within the first 3 days of hospital admission for acute severe UC as conferring a 85% risk of colectomy on that admission.

1.2.2.2 Crohn's disease

Depending on the disease phenotype displayed, different forms of surgical options are available. Patients with perianal disease often need multiple operations to promote abscess drainage and help fistula tract healing. In other cases defunctioning procedures, by diverting faecal flow, can improve inflammation and give extra time for medical therapies to reach therapeutic levels. Finally, many patients, especially those with ileal disease, will require an intestinal resection at some point in their disease course. This is discussed in more detail in Chapter 3.

1.3 Genetics of IBD

The biological pathways underpinning the development of CD and UC are complex and poorly understood. Increasing our understanding of the genetic architecture of IBD will assist scientists in delineating IBD pathways. This will advance our understanding of IBD aetiology and also focus attempts on developing new therapies.

1.3.1 History of the genetics of IBD

1.3.1.1 Twin studies

Although the increased risk of disease development for relatives of IBD patients had already been noted[58], the first definitive evidence of a genetic contribution to the development of IBD was from studies demonstrating a discordance in the incidence of IBD between monozygotic (MZ) and same-sex dizygotic (DZ) twins.[59-61] In these studies, the risk of development of CD in both MZ twins where one had already developed CD was 33-50% whereas in DZ twins it was less than 10%. The genetic contribution to UC development was less as MZ twins showed a concordance of 15-20% compared with about 5% in DZ twins. However, unlike a Mendelian inherited disorder, the concordance was much less than 100% for both diseases in MZ twins, indicating that other determinants (most likely environmental factors) must also have an important role in disease susceptibility.

1.3.1.2 The Early Years

Studies to elucidate genetic susceptibility loci have evolved rapidly in the last 10-20 years with improving technology and knowledge of the human genome. The evolution of technology for genotyping from restriction fragment length polymorphisms (RFLP) to microsatellites to SNPs has catalysed gene discovery. Studies have used either a candidate gene approach, looking at potential genes that have a plausible link with disease pathogenesis and the segregation of markers in genes between patients and their unaffected family members, or a genome-wide approach. Before high resolution sequencing of the human genome had been completed, the genome-wide approach could only identify large segments of the human genome associated with disease in patients compared with their unaffected family members, without being able to identify the actual susceptibility gene(s) within the area. In the 1990s and early 21st century, family linkage studies demonstrated linkage in IBD at various loci in the human genome. They included the *IBD3*[62;63] around the Human Leukocyte Antigen (HLA) Complex on chromosome 6, and *IBD5* containing the *OCTN* gene on chromosome 5.[64] However, narrowing down the signal to a single gene was extremely difficult with the technology available at the time. The only success story was the positional cloning of *NOD2*, which was confirmed independently by two groups to be the CD susceptibility gene at the *IBD1* locus on chromosome 16.[65;66]

1.3.1.3 Genome-wide association studies

Even with improved technology, the ability of family linkage studies to find the common but small-effect polymorphisms that are important in complex diseases like IBD was limited. This was due to the fact that when the effect size is small, some relatives in family studies will be affected because of other causes, and also because of the limited ability to recruit enough families for adequately powered experiments.[67]

With improved resolution of the human genome, genome-wide association studies (GWAS) was the next step, using cases and unrelated controls from the same population and genotyping known SNPs across the genome. Because of linkage disequilibrium (LD) - the tendency of certain areas of the gene to be inherited together - not every variation in the human genome needs to be genotyped. This has heralded a revolution in the genetics of complex diseases. A candidate gene approach requires a hypothesis–driven basis for the selection of genes to be genotyped, whereas GWAS is genome-wide and therefore can be without bias. These studies have significantly progressed the understanding of the biology underlying complex diseases. The first published GWAS in IBD in 2006 genotyped over 300,000 SNPs in 567 patients with ileal CD and a similar number of controls[68]; since then a large number of Caucasian CD susceptibility loci have been uncovered. The most recent international meta-analysis analysed 5 separate GWAS in CD[69-73], and highlighted 71 susceptibility loci.[74] This paper utilised 6333 CD cases and 15056 controls, with a replication cohort of 15694 cases and 14026 controls, and thus had the power to detect susceptibility loci with low odds ratios (OR); the lowest OR detected was 1.04.

Reflecting the lower heritability of UC compared with CD, GWAS in UC did not start for several years after CD GWAS began. McGovern et al[75] did a meta-analysis with two datasets from Los Angeles (USA) and Sweden, and combined them with an existing published study using US-wide cases[76]. A German GWAS[77] and a British GWAS[78] and copy number variation study[79] have been performed. A study in paediatric UC has also been completed.[69] An international UC meta-analysis has recently been completed[80] utilising 6687 cases and 19718 controls, with a replication cohort of 9628 cases and 12917 controls. Overall, 47 loci have a confirmed association with UC susceptibility, conferring OR 1.05-1.74.

Although a lot of work has been done to reveal susceptibility loci, more effort is required: many of the significant tagging SNPs are in intergenic areas, with several plausible genes in the vicinity. Fine mapping is required to identify the candidate gene, and to find the disease susceptibility mutations within the gene. Functional work is then required to identify how changes in the identified gene affect the biological pathways underlying disease.

1.3.2 NOD2

Following the initial genome-wide study in CD in 1996, the first IBD susceptibility locus was identified on chromosome 16, around loci D16S409 and D16S419.[81] Positional cloning identified an association between the candidate gene *NOD2* (nucleotide-binding oligomerization domain containing 2) and CD, the variants being a frameshift mutation (1007fs) and two missense mutations (R702W, G908R) in the leucine rich repeat domain[82] which is thought to be involved in the intracellular sensing of bacterial components.[65;66] Further mutational analyses in 612 patients with IBD confirmed that these three mutations were independently associated with susceptibility to CD[83] and accounted for about 80% of the CD-associated mutations in this gene.

Amongst CD patients the overall prevalence of *NOD2* mutations is around 30-45%[84-86], but there is both ethnic and geographical variation in this prevalence, with an absence of the mutations in Asian populations[87] and a lower contribution to the genetic susceptibility in northern Europe compared with southern European countries.[88] *NOD2* mutations are correlated with susceptibility to ileal disease[84] and stenosing disease.[85]

The ligand of the NOD2 protein is muramyl dipeptide (MDP),[89] which is a component of peptidoglycan present in Gram-positive and Gram-negative bacterial cell walls. The NOD2 protein is located in the cytoplasm and plasma membrane of mammalian cells[90] and appears to act as an intracellular sensor of MDP and therefore of bacteria.[82] NOD2 consists of the LRR domain, involved in bacterial sensing, an

NBD domain and two CARD domains. In the inactive state the LRR domain is folded back on the NBD domain. When a ligand (e.g. MDP) binds to the LRR domain, it induces a conformational change in the NOD2 protein. This allows self-oligomerization between the NBD domains which thereafter causes signal transduction through interaction with RIP2, a CARD containing protein kinase via a CARD-CARD interaction. This complex interacts with NF-κB and causes its activation and relocation to the nucleus to initiate the transcription of inflammatory cytokines (Figure 1-1).

NOD2 is predominantly expressed in Paneth cells in the small bowel[91] as well as monocytes.[92] Paneth cells constitutively produce the antimicrobial peptides α-defensins HD5 and HD6, and reduced defensin production has been shown to occur in ileal CD.[93]

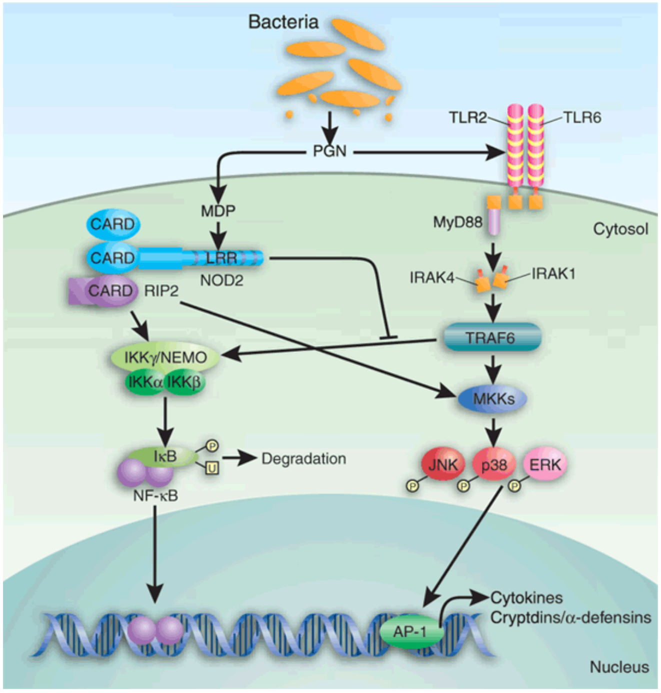


Figure 1-1 Activation and regulation of NOD and TLR pathways by peptidoglycan, from Kelsall *et al.* [94]

The membrane targeting of NOD2 appears to be mediated by 2 leucine residues and a tryptophan-containing motif in the COOH-terminal domain of the protein.[90] Of the three common CD *NOD2* mutations, the 1007fs *NOD2* variant is the only one that is not expressed on the plasma membrane[90] and is unresponsive to MDP.[82] The

R702W and G908R mutations show membrane association, but have reduced levels of NF-κB activation after MDP stimulation compared with the NOD2 wildtype.[90] Even with targeting of the 1007fs to the cell membrane, responsiveness of MDP is not restored.[95] The importance of the membrane localization of NOD2 may be in the recruitment of RICK (RIP2) through a CARD-CARD interaction with subsequent NF-κB activation.[95]

Thus the common NOD2 mutations fit the concept of a defect in innate immunity, which is counter-intuitive in an inflammatory disease. This paradox was addressed in a study[96] which appeared to confirm that there is a weaker immune response to acute insults in patients with CD (though it was not correlated with the *NOD2* genotype), with a secondary exaggerated immune response causing the inflammation. This 'three-stage model' of the imunopathogenesis of CD (initial penetration of foreign material, impaired clearance of this material due to reduced response and then a compensatory adaptive response)[97] has not yet been conclusively proven, but remains an attractive hypothesis.

1.3.3 Autophagy

ATG16L1 is encoded by a gene on chromosome 2q37.1. A non synonymous SNP GWAS[98] first demonstrated an association with the *ATG16L1* SNP rs2241880 and CD susceptibility. The association of *ATG16L1* with CD has been confirmed by several GWAS as well as other genotyping studies, culminating in the international CD GWAS meta-analysis[74] which demonstrated an OR of 1.34 (95% CI 1.29-1.40, $p=6.79 \times 10^{-41}$) for the *ATG16L1* SNP rs3792109, which is in LD with rs2241880. There may also be a weaker association with UC as a recent meta-analysis of all studies published on *ATG16L1* (GWAS as well as individual genotyping studies) suggested an OR of 1.06 (95% CI 1.01-1.1, p=0.02)[99] for UC, although the recent UC GWAS meta-analysis[80] has not shown *ATG16L1* to be associated with UC susceptibility. The functional implications of the threonine to alanine substitution of the rs2241880 (T300A) mutation are not fully understood, although in human epithelial cell lines the alanine substitution impairs capture of internalised *Salmonella* within autophagosomes.[100]

IRGM (immunity-related p47 guanosine triphosphatase family, M) appears to be important in the regulation of autophagy and the elimination of intracellular bacteria.[101] Following the association of SNPs in *IRGM* and CD susceptibility, the recent CD meta-analysis[74] confirmed that the *IRGM* SNP rs7714584 confers an OR of 1.37 (95% CI 1.28-1.47, p-value 7.76 x 10^{-19}) for disease susceptibility. Exonic sequencing of the gene in 248 individuals failed to demonstrate any protein altering mutations.[102] However, the non-functional mutations associated with CD are in LD with a 20kb deletion polymorphism immediately upstream of the gene.[103] This deletion polymorphism is associated with reduced *IRGM* expression, which in turn significantly impairs autophagy in *Salmonella* infected epithelial cell lines.[103]

Autophagy is a well conserved cytoplasmic mechanism which serves both as the way that damaged intracellular organelles are eliminated or recycled and also as a key mechanism for the removal of intracellular bacteria and viruses.[104;105] It is a catabolic process which degrades these items through the lysosomal machinery. There are a number of autophagic processes but the most important one involves the formation of a membrane (the autophagosome) around the targeted area which then fuses with the lysosome.

As shown in Figure 1-2, manufacture of the autophagosome is initiated by starvation or various growth factors. ATG16L1 forms a complex with an ATG5-12 conjugate which drives the autophagocytic process. An LC3 protein complex also promotes the process. The final stages of autophagy, culminating with fusion with the lysosome, are less well characterised but appear to involve LC3.

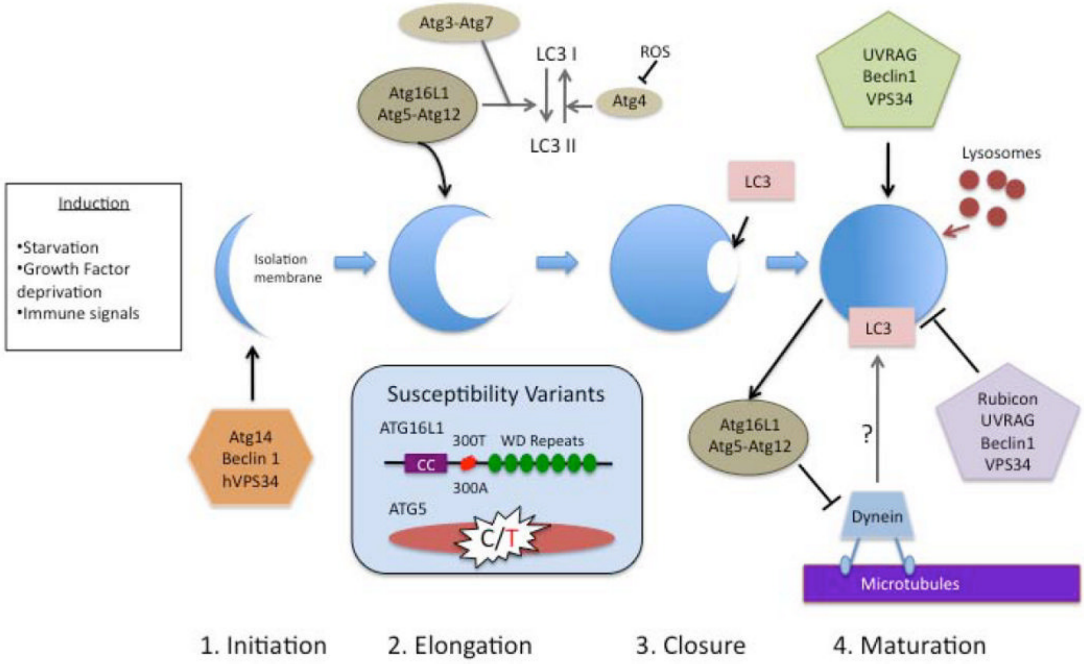


Figure 1-2 The autophagocytic pathway, from Heath and Xavier[106]

Defective removal of intracellular bacteria in the autophagy process provides a plausible link with the pathogenesis of CD, as well as future therapeutic possibilities if autophagy can be upregulated pharmacologically.

Several recent studies have highlighted a functional link between *NOD2* and autophagy. Signalling through toll-like receptors (TLRs) has been shown to lead to autophagosome formation.[107;108] *NOD1* and *NOD2* interact with and recruit *ATG16L1* to the plasma membrane to initiate autophagy, with the 1007fs *NOD2* mutation impairing this recruitment.[109]

1.3.4 Th17 pathway

Conventional wisdom of CD being predominantly a Th1 mediated disease and UC being predominantly a Th2 mediated disease has been re-evaluated in the light of the recent discovery of the Th17 pathway.[110] The Th17 pathway is the production of IL17 from $CD4^+$ T lymphocytes, and is antagonised by key cytokines from the Th1 and Th2 pathways (IFNγ and IL4).[110] IL23 is an important cytokine in the induction of the Th17 pathway. The IL23 cytokine is made up of p19 and p40 subunits, with the p40 subunit being identical to the p40 subunit of IL12. Transgenic mice with ubiquitous expression of IL23 show severe systemic inflammation[111], and IL23 p19 knockout mice show impaired humoral and delayed type hypersensitivity reactions.[112]

Engagement of IL23 in the IL23 receptor complex causes activation of the JAK-STAT signalling pathway, as shown in Figure 1-3. Tyrosine phosphorylation activates STAT which translocates to the nucleus to trigger the expression of cytokines, including IL17A, IL22 and IL21[113]. IL17F and IL22 mRNA expression, induced via IL23 signalling, are increased in inflamed intestinal tissue compared to uninflamed tissue and increased IL22 serum levels have been reported in CD patients[114], suggesting increased activity of this pathway in IBD.

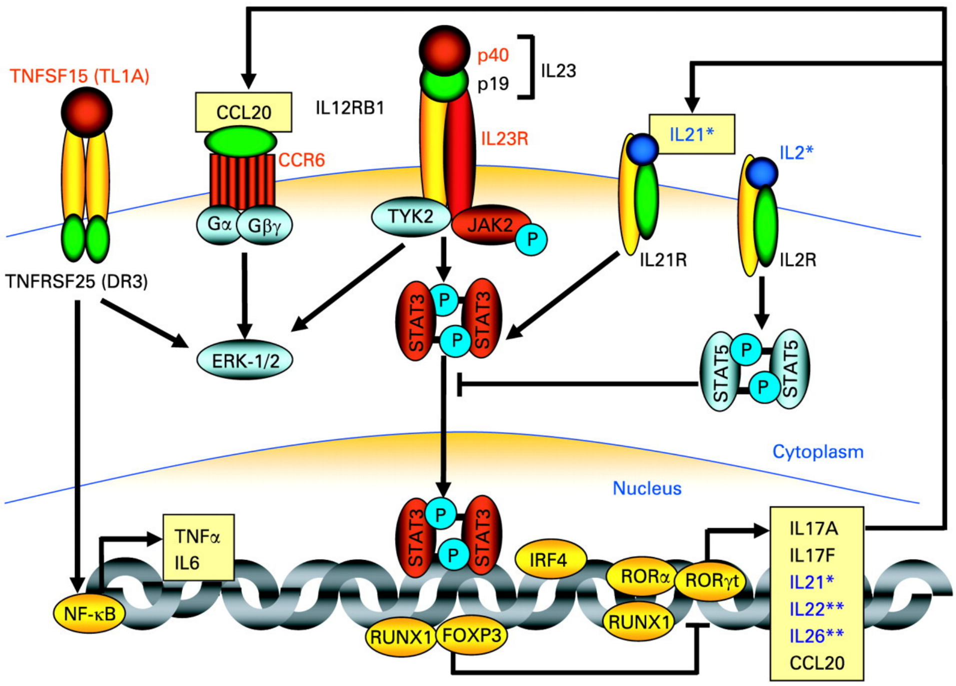


Figure 1-3 Proteins involved in IL23R signalling and Th17 cell development, from Brand[115]

The recent advances in IBD genetics have provided further corroborative evidence of the importance of the Th17 pathway in disease pathogenesis. Multiple genes within the pathway, including *IL23R*, *TNFSF15*, *STAT3*, *IL12B*, *CCR6* and *JAK2* (demonstrated in Figure 1-3) show evidence of a link with IBD susceptibility, as shown in Table 1-2.

Gene/ locus	SNP	Disease	Study	Paper	P-value	OR	95% CI
IL23R	rs11465804	IBD	Paediatric	Imielinski *et al.*[69]	1.03×10^{-15}	0.39	0.29–0.52
IL23R	rs11209026	CD	Meta-analysis	Franke *et al.*[74]	1.00×10^{-64}	2.66	2.36-3.00
IL23R	rs11209026	UC	Meta-analysis	Anderson et al.[80]	5.12×10^{-28}	1.74	1.57-1.92
IL23R	rs11805303	UC	Non-syn SNP	Fisher *et al.*[116]	2.2×10^{-4}	1.24	1.10-1.39
TNFSF15	rs6478108	IBD	Paediatric	Imielinski *et al.*[69]	5.06×10^{-10}	0.74	0.67–0.83
TNFSF15	rs3810936	CD	Meta-analysis	Franke *et al.*[74]	1.00×10^{-15}	1.21	1.15-1.27
TNFSF15	rs4246905	UC	Meta-analysis	Anderson et al.[80]	6.5×10^{-13}	1.13	1.07-1.18
STAT3	rs744166	CD	CD GWAS	Barrett *et al.*[117]	6.82×10^{-12}	1.18	
IL12B	rs6556412	CD	Meta-analysis	Franke *et al.*[74]	4.37×10^{-8}	1.18	1.13-1.24
IL12B	rs6556416	UC	Non-syn SNP	Fisher *et al.*[116]	6.8×10^{-4}	0.84	0.73-0.97
IL12B	rs6871626	UC	Meta-analysis	Anderson *et al.*[80]	4.5×10^{-20}	1.18	1.13-1.24
CCR6	rs415890	CD	Meta-analysis	Franke *et al.*[74]	2.51×10^{-12}	1.17	1.12-1.22
JAK2	rs10758669	IBD	Paediatric	Imielinski *et al.*[69]	5.11×10^{-7}		
JAK2	rs10758669	CD	Meta-analysis	Franke *et al.*[74]	1.00×10^{-13}	1.18	1.13-1.23
JAK2	rs10974914	UC	UC GWAS	Barrett *et al.*[78]	1.5×10^{-5}		

Table 1-2 Synopsis of Th17 genes associated with IBD susceptibility

The best elucidated of these genes and most strongly associated with CD is *IL23R*. It was first associated with CD susceptibility in the first published GWAS in IBD[68], and has since been replicated in other populations. IL23R is expressed most highly on activated T cells, especially Th17 cells as well as natural killer cells.[118] The only functional variant identified by GWAS, rs11209026, confers an OR of 0.45 for CD (minor allele frequency (MAF) 1.9% in CD, 7% in controls) and an OR of 0.55 for UC development (MAF 3.7% in UC, 7% in controls), implying a protective effect in both CD and UC. More functional work is required, but preliminary data suggests that the rs11209026 ‘protective’ mutation reduces STAT phosphorylation in cell lines which are responsive to IL23 (Dr Pidasheva, personal communication), thus preventing normal signalling of the IL23 pathway.

1.3.5 Major histocompatibility complex

In humans the major histocompatibility complex (MHC) is known as the human leukocyte antigen (HLA) complex. The HLA region (Figure 1-4) encodes a number of proteins that are displayed on the outer surface of cells and are important in the differentiation between self and non-self. The region (6p21.1-23) is the most gene-dense area of the genome.

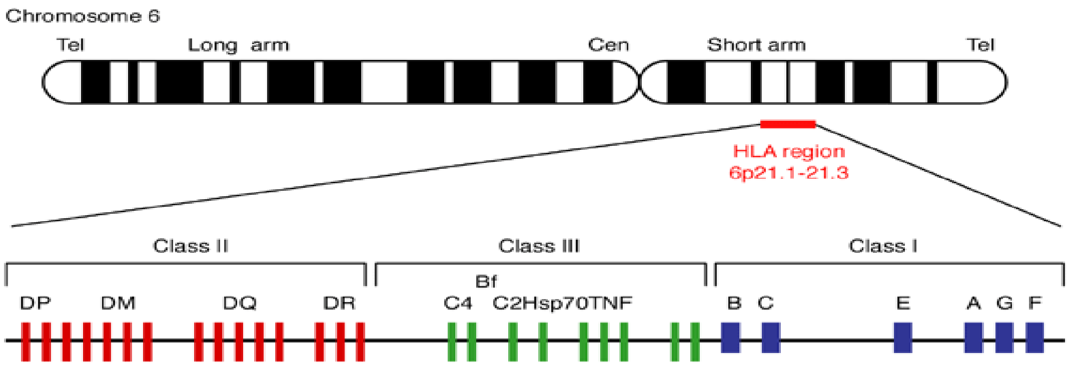


Figure 1-4 Genetic map of the human leukocyte antigen region from Mehra and Kaur[119]

IBD3 was one of the first areas linked with IBD susceptibility and includes the area encoding the major histocompatibility complex. Early studies prior to the days of GWAS linked IBD3 to both CD and UC[62;63], though more so with UC.[120] Most work has concentrated on the Class II MHC, with different alleles providing either risk or protection. A meta-analysis of the relationship between HLA class II phenotypes and IBD in 1999 confirmed this link.[121] Identifying the relevant gene in the area has proved problematic due to the large number of genes and strong LD across the area, and GWAS has not been able to find the causative gene. Multiple alleles within the region have been associated with IBD in recent GWAS, as shown in Table 1-3.

SNP	Postiton	Disease	Study		P-value	OR	95% CI
rs1799964	6p21.33	CD	CD meta	Franke *et al.*[74]	3.98×10^{-11}	1.19	1.13-1.25
rs6908425	6p22.3	CD	CD meta	Franke *et al.*[74]	1.4×10^{-8}	1.17	1.11-1.23
rs943072	6p21.1	UC	UC meta	Anderson *et al.*[80]	1.4×10^{-9}	1.15	1.07 - 1.24
rs9268853	6p21.32	UC	UC meta	Anderson *et al.*[80]	1.35×10^{-55}	1.4	1.34 - 1.47
rs9268480	6p21.32	UC	Non-syn SNP UC	Fisher *et al.*[116]	7.2×10^{-6}		
rs660895	6p21.32	UC	Non-syn SNP UC	Fisher *et al.*[116]	2.8×10^{-8}		
rs477515	6p21.32	IBD	Paed IBD	Kugathasan *et al.*[122]	1.02×10^{-8}	0.724	0.65–0.81
rs9271568	6p21.32	IBD	Paed IBD	Kugathasan *et al.*[122]	2.95×10^{-8}	0.724	0.65–0.81

Table 1-3 IBD3 alleles associated with IBD in recent GWAS

1.3.6 Barrier function susceptibility genes

1.3.6.1 MDR1

The multidrug resistance gene 1(*ABCB1/MDR1*) on chromosome 7q, has been of interest since the first suggestions of association with IBD.[123] A mouse knock-out model (*mdr1a*$^{-/-}$) develops spontaneous colitis when kept in a pathogen free environment and shows dysregulated epithelial growth.[124] The gene encodes for a transmembrane protein (P-glycoprotein 170) that functions as chloride channel pump across epithelial cells, and belongs to the ATP binding cassette superfamily. P-glycoprotein 170 is highly expressed in the human gastrointestinal tract[125] and on the surface of peripheral blood lymphocytes.[126] Two SNPs correlate with P-glycoprotein 170 expression and activity: C3435T and G2677T/A.[127] Conflicting data surround these two SNPs in CD and UC. An association of the C3435T allele with was shown IBD, particularly UC[128;128-130], but was not found in other studies.[131-133] The G2677T/A allele has been shown to be associated with refractory CD.[129] Given that the 2 alleles are in tight LD, 2-locus haplotype association tests have shown increased susceptibility to UC with the 3435T/G2677 haplotype.[130] A more recent meta-analysis[134] showed a significant association of the 3435T allele with UC but not with CD. However, this allele could lie in LD with the causal variant. This hypothesis is supported by a gene wide haplotype tagging approach that has provided robust evidence for the involvement of the *MDR1* gene in susceptibility to UC.[135] However, a later functional study suggested an alternative explanation, as the C3435T polymorphism, despite being 'silent', seems to affect P-glycoprotein function and substrate specificity.[136]

1.3.6.2 DLG5

DLG5 is on chromosome 10q23. The DLG5 protein is a member of the membrane associated guanylate kinase family that appears to be involved in protein-protein interactions and is likely to be a scaffold protein involved in the maintenance of epithelial integrity.[137]

Positional cloning identified genetic variants in the gene associated with IBD in German and British populations.[138] The *DLG5* gene is in an area of strong LD, and the extended *DLG5* haplotype has 4 common haplotypes with haplotype A being under-transmitted and haplotype D being over-transmitted in IBD. One of the main SNPs in haplotype D is the SNP 113G→A, causing an amino acid substitution R30Q, which may disrupt binding to a Rab GTPase and therefore is likely to have functional implications. Though this finding has been replicated in Canadian and British populations[139], and in a Scottish early-onset population[140], it has not been replicated in other cohorts: Scottish adults[141], German [142] and English patients.[143;144] In addition, a meta-analysis of published studies failed to demonstrate an association.[145] This variability in association may be due to population heterogeneity, gender or gene-gene interactions.

1.3.7 Other genes

Recent GWAS have implicated multiple other SNPs in IBD susceptibility. However, as most of these SNPs are in intergenic areas, defining the exact gene responsible for the association is not straightforward.

1.3.7.1 IBD5

It was a genome wide scan that suggested linkage in the 5q31-33 region, with higher density mapping further pinpointing a locus contributing to CD susceptibility in patients with early onset disease.[63] This association has been confirmed in multiple independent panels of patients.[146-149] Within this area a 250Kbp haplotype has been shown to be associated with CD.[150] However, localising the causal variant has proved problematic, due to tight LD and the large number of candidate genes in the area. One of the candidate genes is *OCTN1/2*. This was first reported in a study where two variants (a missense substitution in *SLC22A4* and a G to C transversion in the *SLC22A5* promoter) in the *OCTN* cluster (now known as *OCTN1* and *OCTN2* respectively) formed a haplotype associated with susceptibility to CD in Canadian patients of European origin, independently of the full *IBD5* haplotype.[151] However, further studies amongst Scottish patients[152], Caucasian children[153] and North American patients[154] found that the association with *OCTN1/2* and susceptibility to CD was only in the context of the IBD5 risk haplotype. Even more conflicting was data from a Belgian cohort that found no association between the *OCTN1/2* variants and susceptibility to CD.[155] Controversy continues as to the phenotype displayed by the *IBD5* genotype and whether there is epistasis between *IBD5* and *NOD2*.

1.3.8 Missing heritability

Even with the current 71 CD susceptibility loci, it has been stated that only 23.2% of the heritability of CD can be explained.[74] However, the majority of the loci uncovered by GWAS are unlikely to be the causal allele for disease susceptibility. It has therefore been argued that as 'imperfect proxies' for the causal allele, the true heritability at any loci is underestimated, so in fact the proportion of heritability explained by these 71 loci may be much higher.[74] In addition, current heritability estimates may not be accurate as new updated twin data suggests lower disease concurrence than previously thought.[156] Twin studies also use the 'equal environments assumption': that dizygotic twins have an equally shared environment compared with monozygotic twins. If this is not the case, then heritability will be overestimated.

It is a matter of debate whether any remaining heritability is explained by many thousands of alleles of very marginal OR, or whether, by pooling populations, we are diluting out population-specific loci with much higher OR. Population mixing is necessitated by GWAS studies in order to have sufficient power for genome-wide statistical significance. As the SNPs that are genotyped in GWAS are present in 5%

or more of the population, other SNPs of lower prevalence may also be important in disease susceptibility. In addition, there are likely to be many 'private mutations' that are family-specific and may involve deletions, duplications and inversions. Structural variation, for example copy number variation (CNV), may also contribute to disease susceptibility, something that is only beginning to be explored.

As next generation sequencing technologies are now becoming cheaper, it will soon be possible to sequence individuals, facilitated by the 1000 genomes project which is providing a comprehensive list of variants with MAF $\geq$1%. This is the most likely method whereby rare SNPs will be identified. However, identifying causal alleles will be difficult due to the large volume of data that will be generated.

1.4 Environmental factors in IBD

There is increasing recognition of the role of environmental factors in changing genetic phenotype without causing germline genetic mutations (termed 'epigenetics'). This phenomenon uses various mechanisms including DNA methylation and histone acetylation to change gene expression. It is possible that some of the environmental factors in IBD may act through epigenetic mechanisms; this is likely to be an increasingly important area for research in the future.

1.4.1 Smoking

Smoking is known to increase the risk of developing CD, and reduce the risk of UC. A recent meta-analysis[157] comparing current and never smokers demonstrated an OR of 1.79 (95% CI 1.4-2.22) and 0.58 (95% CI 0.45-0.75) for development of CD and UC development. A study looking at sibling pairs discordant for both smoking and IBD type (UC or CD) showed that smokers developed CD and non-smokers developed UC, suggesting a gene-environment interaction contributing to disease development in genetically susceptible people.[158]

Smoking appears to affect the disease course for both CD and UC. In CD, smokers tend to develop small bowel rather than colonic disease.[159] There is evidence that those who quit during their disease have a lower risk of subsequent flare-up than those who continue smoking.[160] With UC, there are anecdotal reports of patients who have flare-ups on attempting to stop smoking, with improvement of disease activity course when they start to smoke again, but this association has not been proven conclusively.

The active ingredient in cigarette smoke is unknown. With over 4000 compounds, many of which are carcinogenic, definitive studies to find the agent are unlikely in the near future. Studies trialling nicotine transdermal patches showed no benefit in the maintenance of remission in UC[161], but a modest benefit in active disease.[162] An open label pilot study of nicotine enemas in colonic CD suggested some benefit[163], although a randomised trial in active UC failed to show any benefit.[164]

1.4.2 Previous appendectomy

Previous appendectomy reduces the subsequent risk of UC development (OR 0.307 (95% CI 0.249-0.377).[165] For CD there appears to be an increased risk of disease development following appendectomy[166], although as the increased risk is predominantly in the first year after appendectomy it is unclear whether the symptoms that led to appendectomy were actually unrecognised symptoms of CD.

1.4.3 Diet

Ongoing controversy is provided by the role of diet in disease susceptibility, and further studies are required in this area. The importance of diet in the modulation of CD is demonstrated by the fact that exclusive enteral feeding can induce remission in active disease (reviewed by O'Sullivan[167]). In a Japanese case-control cohort with both UC and CD patients, the risk of IBD development was higher with increased refined sugar consumption, although this was a retrospective study and may suffer from recall bias.[168] In North American children a case-control study demonstrated that low fruit and vegetable consumption was associated with CD development.[169] High levels of arachidonic acid, present in red meat as well as some oils and margarines have been found in the adipose tissues of patients who subsequently went on to develop UC compared to those that did not[170], and increased consumption of arachidonic acid is linked with an increased incidence of the disease.[171]

1.4.4 Hygiene hypothesis

The hygiene hypothesis, first formally proposed in 1989[172], is the theory that lack of early exposure to microbes leads to an increased susceptibility to allergic diseases later in life. This could explain the increasing incidence of allergic diseases in the Western world, and their continued low incidence in the developing world. However whether CD and UC also fit the hygiene hypothesis is unclear. Certainly, observational studies have linked IBD susceptibility with affluence[173] although not with urban living.[174]

1.4.5 Microbes

The gastrointestinal tract is the most heavily colonised area of the human body with the number of bacteria estimated at 10^{14} composed of at least 400 species. Commensal bacterial communities in the gut have a symbiotic relationship with the host, helping to metabolise otherwise indigestible complex carbohydrates, challenging and promoting immune system development and regulating gut development. These bacteria have evolved metabolic mechanisms to acquire nutrition from the host diet, unlike pathogenic bacteria which have virulence factors to enable them to access the tissues of the host.[175] The large number of symbiotic bacteria make it more difficult for pathogenic bacteria to gain access to the host. Whether the host immune system is tolerant or ignorant of symbiotic bacteria is unclear, but it would appear that commensals reside in the lumen and the outer mucus layer, rather than the deep mucus layers and epithelial cells. Evidence would suggest that minimising host recognition of commensals by keeping them within the intestinal microenvironment is important. Innate and adaptive immunity complement each other in protecting the host, as shown by work in MyD88$^{-/-}$ Ticam$^{-/-}$ mice (deficient in the toll-like receptor adaptor molecules MyD88 and Ticam, important in innate immunity) reared in germ-free environments but then challenged with commensal bacteria. These mice produce high titres of serum antibodies against commensal bacteria unlike their control wild type counterparts, who are able to maintain the bacteria within the mucosal immune compartment.[176] The mechanisms underlying this are unclear, but it does not appear to be due solely to increased intestinal permeability.[176] Certainly it would appear that abnormal host responses, in particular in innate immunity, may contribute to bacteria gaining access to compartments beyond the gut mucosa. This could be particularly pertinent in

IBD, where defects in innate immunity appear to be important in disease pathogenesis (see sections 1.3.2 and 1.3.3).

There are many reasons to suspect microbial involvement in disease development and/or perpetuation. As already described, many of the susceptibility genes in IBD are important in innate and adaptive immunity. In animal models, disease does not occur in animals reared in a germ-free environment, unlike those that have been colonized by bacteria.[177] Prolonged courses of antibiotics can improve CD symptoms in some patients.[178] Diversion of faecal stream has been shown to improve inflammation in CD, with re-exposure to luminal contents in an excluded ileum causing recurrent inflammation.[179] In addition, there is evidence to suggest that an episode of infectious gastroenteritis increases subsequent risk of IBD.[180;181]

Whether microbe involvement in IBD pathogenesis is due to an abnormal immune response to gut commensals, or whether there are specific pathogens is unclear; although several microbes have been implicated in IBD, none have been conclusively proved to be causative in disease development.

1.4.5.1 Adherent-invasive Escherichia coli

Escherichia coli is a Gram-negative rod and many strains are important commensal bacteria in the gastrointestinal tract, colonizing the neonatal gut within hours of birth and helping to maintain intestinal homeostasis; other strains are recognised to cause gastroenteritis, usually through food poisoning.[182] Adherent-invasive *Escherichia coli* (AIEC), unlike other pathogenic strains of *E. coli*, lack conventional genes conferring invasiveness, and have been isolated from the mucosa of patients with ileal CD undergoing resection[183] and CD biopsy samples.[184] *E. coli* antibody titres are higher in CD patients than in healthy controls[185] and high numbers (37-56%) of CD patients have antibodies to OmpC (*E. coli* outer membrane porin C).[186-188] OmpC is thought to play a critical role in bacterial adhesion and invasion.[189] AIEC have been shown to survive and replicate within macrophages, producing large amounts of TNFα.[190] This is thought to be secondary to impaired autophagy, a key mechanism in the clearance of intracellular bacteria[191], thus linking key CD susceptibility genes directly to a bacterial aetiology for CD.

1.4.5.2 Mycobacteria

The similarity between the granulomas of CD and *Mycobacterium species* has long been recognised. Johne's disease, caused by *Mycobacterium avium subspecies paratuberculosis* (*MAP*) has manifestations similar to CD in ruminants: predominant ileal disease, weight loss and diarrhoea.[192] When Chiodini and colleagues cultured *MAP* species from 3 patients with CD[193], scientific plausibility was given to the controversial theory that *MAP* may be the cause of CD in humans. *MAP* has been detected in milk[194] and water supplies.[195] Further evidence has been provided by various studies detecting *MAP*-specific DNA in CD tissues by PCR[196;197] and *in-situ* hybridization[198], although more recently *MAP*-specific DNA was not found in CD patients.[199] However, CD improves rather than deteriorates with immunosuppression, and whilst *MAP* has been detected more frequently in CD patients than controls, it is unclear if it causes disease in a subset of patients, or whether it has a propensity to selectively colonise the diseased gastrointestinal tract of patients who have already developed CD, thus acting as an innocent bystander. Moreover, a double blind placebo controlled trial of two years of therapy with clarithromysin, rifabutin and clofazamine (to eliminate MAP) in CD patients with active disease failed to demonstrate a sustained benefit.[200]

Leprosy is a chronic granulomatous disease caused by *Mycobacterium leprae* and *Mycobacterium lepromatosis*. A GWAS of leprosy susceptibility in a Chinese population, comparing patients with leprosy to controls who had not developed disease despite living in the same area, demonstrated a number of susceptibility loci shared with CD.[201] This confirms that susceptibility to some microbial infections can be affected by a person's genetics. As there are similarities in the pathology of some leprosy cases and CD, it remains possible that *Mycobacteria* species, either *MAP* or a strain as yet unidentified, could be responsible for at least a subset of CD.

1.4.5.3 Faecalibacterium prausnitzii

Intriguing data have emerged of the association of the Firmicute *Faecalibacterium prausnitzii* with CD: a reduced presence in the gut microbiota has been found to be associated particularly with post operative and endoscopic recurrence of ileal disease.[202] This study also found that *F. prausnitzii* has anti inflammatory effects in cellular and mouse models. A further study has shown that UC patients also have reduced *F. prausnitzii* counts.[203]

1.4.5.4 Other microbiota - the human microbiome

Many of the human intestinal commensal bacteria have not been isolated and cultured.[204] Various methods, including 16S rRNA technology, have been used to detect and identify unculturable organisms in the human gastrointestinal tract. Patients with CD have been shown to have a reduced diversity of faecal microbiota[205], and a microarray analysis has demonstrated reduced abundance of *Bacteroides sp.* and increased *Enterococcus sp*. in CD patients compared with controls.[206] Investigating the human microbiome is an exciting future prospect in IBD.

1.5 Barrier protection

Mucosal integrity is important for protecting the gut against intraluminal material, especially microbes. In the three stage model of the immunopathogenesis of CD referred to in section 1.3.2, the initial beginning of the process of disease development is penetration of foreign material into the bowel wall. Thus barrier protection is a key area for investigation in IBD.

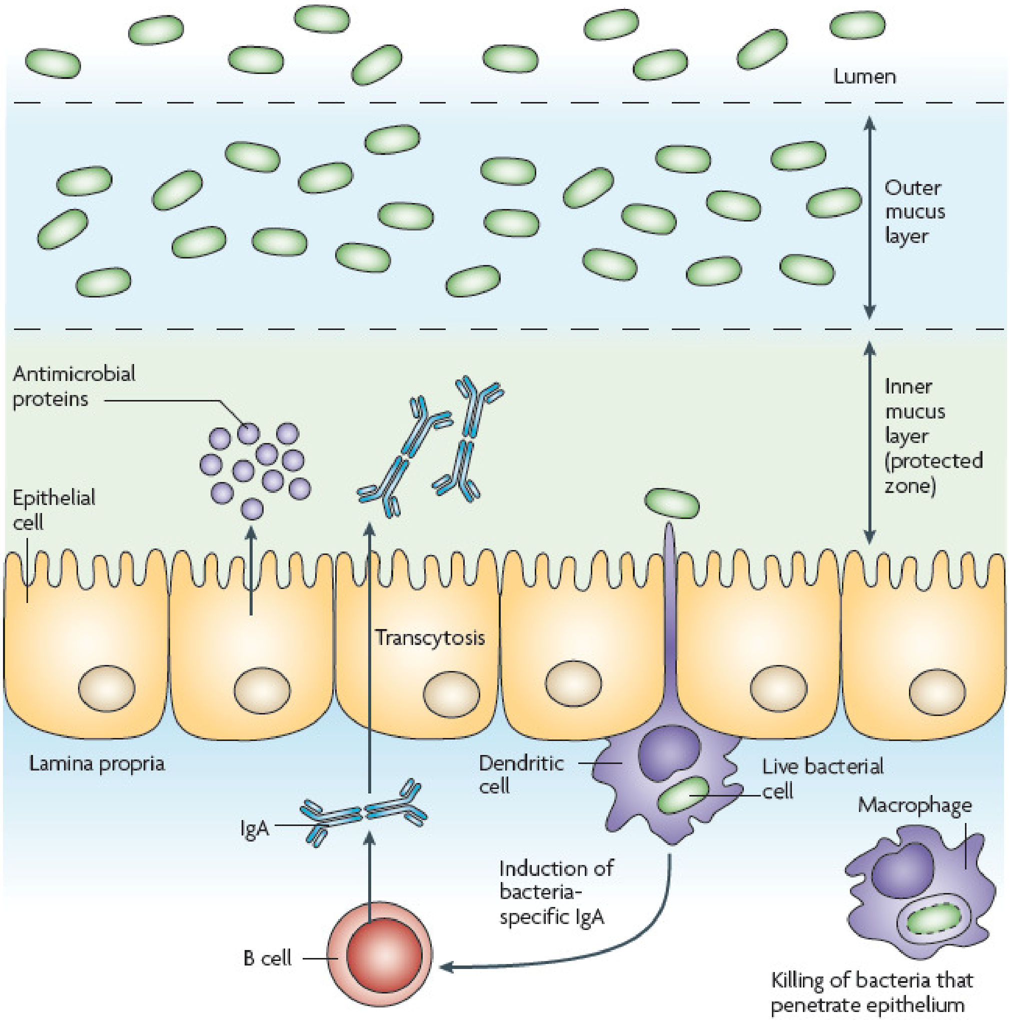
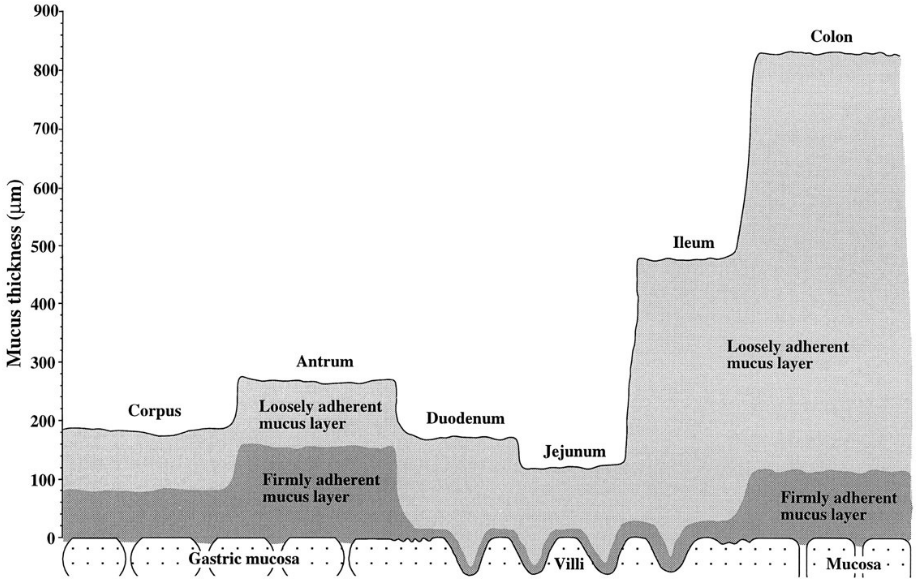


Figure 1-5 Barrier protection of the gastrointestinal tract, from Hooper[207]

Appropriate protection is important to allow essential functions like nutrient and water absorption whilst preventing penetration by luminal pathogens. There are three layers that serve this purpose in the gastrointestinal tract, as shown in Figure 1-5. The outermost layer is a secreted loosely bound mucus gel layer which traps bacteria and pathogens, and contains commensal bacteria. This layer is composed of secretory mucins produced by goblet cells. As shown in Figure 1-6, the thickness of this layer varies through the gastrointestinal tract, from 715µm in the colon to 108µm in the jejunum in rats.[208] Beneath that there is a mucus layer more tightly bound to the luminal cell surface, largely devoid of bacteria, termed the glycocalyx. It consists predominantly of membrane-bound mucins, and also contains immunoglobulins IgA and antimicrobial peptides, e.g. defensins. The depth of this layer in rats varies between minimal or none in the small intestine to 154µm in the gastric antrum.[208]



Mean mucus thickness (µm)						
Total	189 ± 11	274 ± 41	170 ± 38	123 ± 4	480 ± 47	830 ± 110
Firmly Adherent	80 ± 5	154 ± 16	16 ± 3	15 ± 2	29 ± 8	116 ± 51
Loosely Adherent	109 ± 12	120 ± 38	154 ± 39	108 ± 5	447 ± 47	714 ± 109
n	6	6	7	6	5 (8 firm)	11

Figure 1-6 A schematic figure showing the thicknesses of the 2 mucus gel layers *in vivo* in the rat gastrointestinal tract, from Atuma *et al.*[208]

The epithelial cell layer of the mucosa forms the deepest layer providing barrier protection. Epithelial cells, including intestinal enterocytes, secretory Goblet cells, Paneth cells and enteroendocrine cells are all connected to each other by epithelial tight junctions. There is evidence to suggest that CD patients and their unaffected first degree relatives have increased intestinal permeability.[209]

1.5.1 Mucins

Mucins, as already mentioned, can be either membrane bound or secretory. They are heavily glycosylated molecules, resistant to proteases as well as having high water holding capabilities. They are secreted as glycoprotein aggregates of very large molecular masses with the individual molecules linked to one another by non-covalent interactions. By being heavily glycosylated they form the major part of secretions to protect epithelial surfaces and thus appear to be important in the maintenance of mucosal integrity. The secretory mucin MUC2 and the membrane bound MUC3 are the commonest mucins expressed in the gastrointestinal tract.[210]

1.6 Thesis plan

The first aim of the thesis was to define the Scottish CD phenotype according to the Montreal classification, the results of which are presented in Chapter 3. The newly recruited Dundee CD cohort is compared with the more established Edinburgh CD cohort. Characteristics affecting time to development of disease progression (stricturing and/or internally penetrating disease) and time to first resection are examined in univariate and multivariate analyses of the Scottish cohort. Multiple resection data on the Scottish cohort are explored including time between resection with respect to disease location. The accepted convention that CD progresses from inflammatory to stricturing/penetrating, or inflammatory to stricturing to penetrating is examined in the Dundee cohort. Finally, the development of disease progression, time to first resection and multiple resection data results in the Scottish cohort are compared to other published cohorts.

In Chapter 4, the problems with existing definitions of severe CD are explored. A novel score for defining severe CD is introduced, and the results of its application to the Dundee cohort presented. It is compared to the most widely accepted current definition of disease severity. A comparison of clinical and genetic characteristics present at diagnosis between more severe and less severe CD patients is shown, including a model for calculating the probability of more severe disease that could be used at diagnosis. The results of a case-control analysis of 32 CD susceptibility SNPs are also presented.

Having discussed clinical and genetic factors affecting the CD phenotype, the thesis moves on to discuss GALNT2, one of the NOD2-interacting proteins uncovered by the yeast two-hybrid screen. Its role in Scottish IBD susceptibility is explored in Chapter 5, including sub phenotypic analyses in both CD and UC. The results of sequencing the exons of *GALNT2* are also presented in the search for a potentially causative mutation.

Further, more functional, analyses of the interaction between GALNT2 and NOD2 are presented in Chapter 6, including the examination of the level of the interaction, and the examination of how the most common CD NOD2 variants affect the interaction between the two proteins. The results of GALNT2 and NOD2 expression analyses by both immunohistochemistry and quantitative PCR are presented.

In Chapter 7, a case-control susceptibility analysis of the *MUC2* and *MUC3* genes, the commonest mucins in the gastrointestinal tract, is presented. Chapter 8 presents the results of the search in the Scottish IBD cohort for the gene representing the CD GWAS meta-analysis signal on chromosome 12q12 with the results of haplotype-tagging studies of both *MUC19* and *LRRK2*.

Finally, in Chapter 9, the overall implications of the work presented in the thesis are discussed and suggestions made for future research.

Chapter 2 Materials and Methods

2.1 Reagents

All reagents were from Sigma-Aldrich unless otherwise stated.

2.2 Patients

2.2.1 Dundee ethics

The study had the full ethics approval of the Tayside Local Research Ethics Committee (LREC), approval number 226/02.

2.2.2 Dundee recruitment

Inclusion critera: Consenting adult patients of >18 years old with a proven diagnosis of IBD. Patients attending the weekly IBD clinic in Ninewells Hospital were sent a letter by Mrs Shirley Cleary RN at least 1 week prior to clinic appointment inviting them to participate (Appendices 10.1 and 10.2). At clinic, these patients were asked whether they wished to participate and signed a consent form to formally document their agreement to the terms of the study. Patients filled in a questionnaire (Appendix 10.3) and a 15ml sample of blood was obtained in EDTA tubes, or a 5ml sample of saliva into an Oragene® saliva kit (DNA Genotek, Ontario, Canada) for DNA extraction. This clinic recruitment was done with Mrs Shirley Cleary RN, recruiting equal numbers of patients each.

2.2.3 Edinburgh ethics and recruitment

Patients living in the Edinburgh area were recruited at the Western General Hospital. This study had the full approval of the Lothian LREC (2000/4/192). These patients were recruited by Mrs Linda Smith and Mrs Hazel Drummond in the same manner as described for the Dundee recruitment.

2.2.4 Controls

Controls were obtained from several sources. Scottish Central and Tayside region controls for the Dundee cohort were obtained from the Generation Scotland 3D project, as previously described.[211] Controls for the Edinburgh cohort were locally recruited controls who were friends and non-blood relatives of the Edinburgh IBD cohort (recruited by Mrs Linda Smith, RN); additional controls were obtained from the 1958 birth cohort.

2.2.5 Phenotyping

Following clinic attendance, each patient's notes were scrutinised retrospectively to extract details about their disease phenotype, which was documented on a paper proforma developed by Mrs Hazel Drummond (Appendices 10.5 and 10.6). All phenotyping of the Dundee patients was done by myself whereas the Edinburgh patients were phenotyped by Mrs Hazel Drummond and other medical professionals and scientists. The initial diagnosis of IBD was according to the Lennard-Jones criteria[2] (Table 2-1), and subsequent disease location and behaviour were documented carefully according to the Montreal classification[20] (Table 2-2 and Table 2-3). By convention, the CD location was taken to be the maximum extent of macroscopic disease prior to first resection. For patients who had not had a resection by the time of phenotyping, it was the maximum extent during follow up. CD behaviour was taken to be the disease behaviour at 5 years after diagnosis, or at last diagnosis for those who had not yet been diagnosed for 5 years. The UC disease extent was the maximum extent of macroscopic disease at any point during follow up. Patients with IBD-U (unclassified IBD) were included for overall IBD analyses but excluded from CD and UC analyses.

Ulcerative colitis criteria	
Exclusion	Infective colitis Ischaemic colitis Irradiation colitis Solitary ulcer Abnormalities suggesting CD Complex anal lesion Granulomata
Inclusion	Continuous mucoscal inflammation without granulomata Affecting the rectum and some or all of the colon in continuity with the rectum
Crohn's disease criteria	
Exclusion	Infections Ischaemia Irradiation Lymphoma/carcinoma
Inclusion	Mouth to anus: Chronic granulomatous lesion of the lip or buccal mucosa Pyelo-duodenal disease Small bowel disease Chronic anal lesion Discontinuous Transmural: Fissuring ulcers Abscess Fistula Fibrosis Stricture Lymphoid aggregates present Mucin retention in the presence of active inflammation Non caseating granulomata

Table 2-1 Lennard-Jones criteria for IBD diagnosis

Age at diagnosis	A1	<17 years
	A2	17-40 years
	A3	>40 years
Disease location	L1	Ileal +/- caecal overspill
	L2	Colonic
	L3	Ileocolonic
	L4	UGI (+ L1-L3 if both UGI disease and elsewhere)
Disease behaviour	B1	Inflammatory
	B2	Stricturing
	B3	Internally penetrating
	p	Perianal modifier

Table 2-2 Montreal classification – Crohn’s disease

E1	Rectal
E2	Distal to splenic flexure
E3	Proximal to splenic flexure

Table 2-3 Montreal classification - Ulcerative colitis

Additional information was also recorded for each CD patient including drug history, number of hospital admissions and BMI in the first 5 years in order to calculate a severity score (Table 2-4). Further details of the development of this score are given in section 4.2.1.

2.2.6 Databases

A custom-designed Microsoft Access™ database was used for the storage of the phenotypic information from the paper proforma on computer. This database had been developed some years ago in Edinburgh and was used for the storage of data from the Edinburgh cohort. It was adapted with the help of Mrs Maureen Edwards (Database Manager, Clinical Genetics, Western General Hospital, Edinburgh) for the Dundee IBD cohort. All phenotypic information was entered onto this database.

	Severe: Score 4 for each	Moderate: Score 2 for each	Mild: Score 1 for each	Score 0
Disease extent/behaviour	**Panenteric disease** **OR Complex perianal disease requiring 3 or more operations** **OR Fistulating disease**	**Stricturing but not fistulation** **Perianal disease requiring 1 or 2 operations**	**Single site involvement** **No evidence of stricturing/fistulation** **Perianal disease but not needing operation**	
Medical/surgical management	**Steroid dependency** **OR Need for 2 or more immunomodulatory drugs** **OR 2 Surgical resections** **OR Use of biological therapy**	**More than 4 steroid courses, but none >4 months** **OR 1 immunomodulator** **OR 1 surgical resection**	**1-3 courses of steroids, each lasting <4 months** **No immunomodulators**	**No steroids** **No immunomodulators**
Nutritional status	**BMI <15 at any point in the 5 years**	**BMI 15-18.5 at any point in the 5 years**		**BMI >18.5 at all times in the 5 years**
Socio-economic impact	**5 or more hospitalizations for management of active disease**	**2-4 hospitalizations for active disease**	**1 hospitalization for active disease**	**No hospitalizations for active disease**

Table 2-4 Severity score, calculated for the first 5 years after diagnosis

2.3 DNA

2.3.1 DNA extraction

DNA extraction from blood and saliva was outsourced to the Medical Research Council (MRC) Human Genetics Unit, Western General Hospital, Edinburgh. They used the Nucleon DNA extraction kit (Gen-Probe, Manchester, UK), using chloroform and Nucleon resin for DNA extraction and ethanol for DNA precipitation. DNA concentrations were quantified at the MRC by Picogreen® (Molecular Probes, Oregon, USA).

In the GI Unit the sample concentrations were confirmed using a Nanodrop 1000 machine (Thermo Sciences, UK) and each sample made up to 150ng/µl using 1xTE (10mM Tris pH8 and 1mM EDTA). Those with initial concentrations significantly below this value were re-precipitated and resuspended in a smaller volume to concentrate them. DNA samples were stored at -80°C.

2.3.2 SNP selection

SNPs were selected in two different ways. Tagging SNPs were chosen for a gene of interest to perform fine mapping of genetic variation. To choose tagging SNPs for a gene of interest, the genetic variation for the gene was downloaded from www.hapmap.com[212], allowing for 15kbp at either end of the gene. This information was run in Haploview 4.1 and haplotype blocks defined according to solid spine of linkage disequilibrium. The SNPs were chosen to tag haplotypic variation (>5%) of the gene of interest. In other experiments SNPs were chosen to try to replicate other people's results.

2.3.3 Taqman® genotyping

Taqman® genotyping (Applied Biosystems, Life Technologies) was completed at the Wellcome Trust Clinical Research Facility, Western General Hospital, Edinburgh on the ABI PRISM 7900HT machine. Taqman® genotyping involved the design of nucleotide probes – one each for each of the variants of the SNP being examined. Each probe was tagged at the 5' end with a different fluorophore and had a quencher at the 3' end. When still in close proximity to the quencher, the fluorophore was prevented from emitting any fluorescence. During real time polymerase chain reaction (RT-PCR) amplification of the target sequence, when the probe annealed to the genomic DNA, the 5’-3’ exonuclease activity of the DNA Taq Polymerase degraded the probe, thus releasing the fluorophore from the blocking of the quencher. The amplification conditions were: 50°C for 2 minutes, 95°C for 10 minutes then 40 cycles of 95°C for 15 seconds and 60°C for 1 minute. As both probes had different fluorophores attached to them, comparing the fluorescence of the emissions from the samples showed clustering according to the genotype of the SNP in question (AA, AB, BB). This was visualised on an allelic discrimination assay plot where the fluorophores were plotted on opposing axes (Figure 2-1). In circumstances where the genotype could be confidently called the sample was dropped from the analysis.

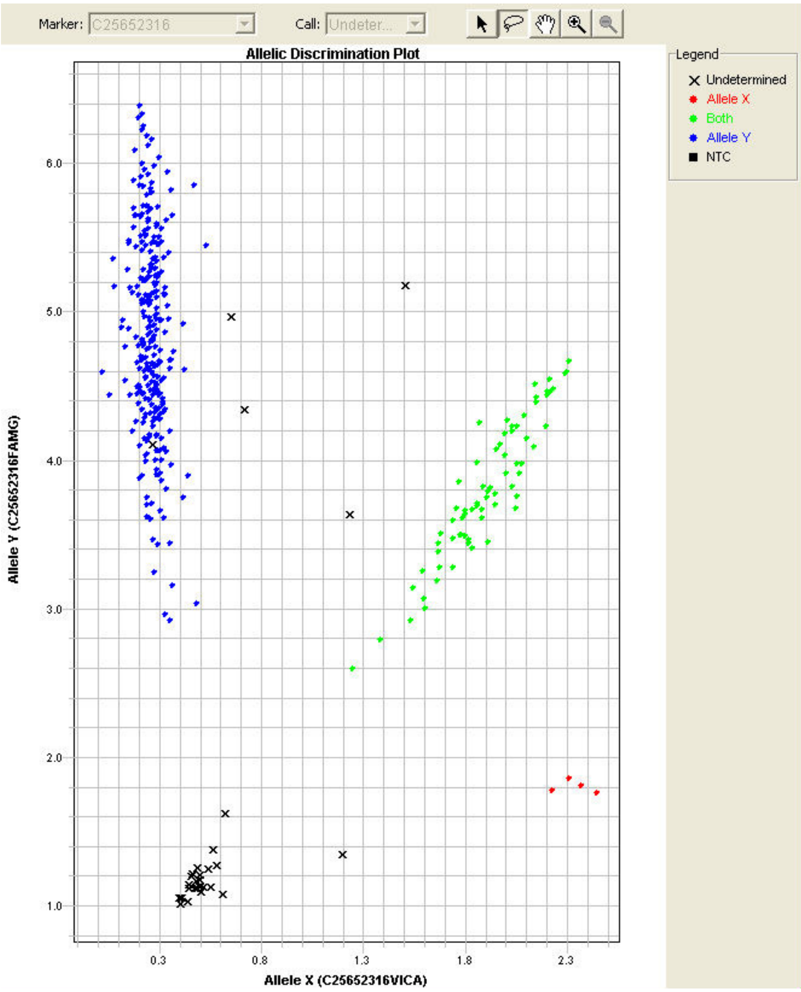


Figure 2-1 Example of Taqman® clustering

2.3.4 Sequenom® genotyping

Sequenom® sequencing was completed at the Genomics Core of the University of California, San Francisco. Rather than using fluoroscent probes, SNPs were detected using MALDI-TOF to calculate the mass of the PCR extension product and hence the genotype of the SNP in question.

2.3.5 Illumina Goldengate® genotyping

With the Goldengate® platform (Illumina, SanDiego, CA), three oligonucleotides were designed for each SNP: two allele specific oligonucleotides (ASO) and a downstream locus specific oligonucleotide (LSO). All three of these also had sequences complementary to universal primers – P1, P2 and P3, with each allele of the SNP of interest being specific to P1 or P2. The oligonucleotides were allowed to hybridise to the genomic DNA being analysed. Extension of the ASO to the LSO joined information about the SNP allele present and the SNP location. The three universal primers were then used to amplify the product. P1 and P2 were Cy3- and Cy5-labelled respectively, and the dye-labelled PCR products were allowed to anneal to a bead containing a complementary sequence to the LSO. Thus, by analysing the fluorescence of that bead, the genotype (AA, AB or BB) was determined in a similar way to Taqman®. As a chip could contain many beads and the universal primers were not specific to each reaction, many SNPs could be multiplexed.

2.3.6 Data quality control

For all three methods of genotyping, similar methods for ensuring the quality of the data were employed. Initially the control samples were analysed using Haploview and samples not in Hardy-Weinberg (HW) equilibrium were discarded (HW p-value<0.05, corrected appropriately for multiple testing). SNPs with less than 90% genotyping were also removed from the analysis. Where multiple SNPs were investigated (eg the work using the Illumina and Sequenom®), individual DNAs with <90% successful calling on the remaining SNPs were not used in analyses.

2.3.7 Analysis of genotyping

SNP genotyping was analysed using Haploview version 4.2. For tagging SNP data, haplotypes were defined and analysis performed according to the initial haplotypes on which the SNPs had been selected. P-values of <0.05 were considered significant, and corrected for multiple testing as appropriate.

2.3.8 Polymerase Chain Reaction (PCR)

The sequence of the area of interest was downloaded from www.ensembl.org. Pairs of primers were designed to cover the relevant SNP according to the criteria: 20-22 base pairs long, an approximate GC content of 50%, the 3' end being a G or a C, and not overlying a known SNP. The primers were selected to be at least 60 base pairs away from the 5' and 3' of the sequence to be analysed. Using the Sigma website (www.sigmaaldrich.com) the sequences were checked for the melting temperature of the primer, aiming for about 60°C and preferably for the absence of a secondary structure.

For the PCR, each primer was used at a working concentration of 1mM, with NH_4 buffer (Bioline, London, UK, containing at working concentration 16mM $(NH_4)_2SO_4$, 67mM Tris HCl at pH 8.8 and 0.01% Tween 20), $MgCl_2$ (working concentration 1.5mM), dNTP (Bioline, London, UK, working concentration 1mM) and 0.2µl Taq polymerase (working concentration 0.05U/µl, gift of Dr Elaine Nimmo), and the appropriate DNA (final concentration 2.5-5ng/µl).

The optimal temperature at which to run the PCR was checked using a temperature gradient range of 55-65°C, as shown in Table 2-5.

94°C for 2 minutes (initial denaturation)	
92°C for 30 sec (denaturation)	}
Target temp for 30 sec (annealing)	} repeated
72°C for 45 sec (extension)	} 31 more times
72°C for 5 minutes to complete the PCR	

Table 2-5 PCR conditions

2.3.9 PCR product gel electrophoresis

The samples were run on a 1% agarose gel (agarose in 0.5xTBE) . The 10xTBE stock solution used contained 106g Tris base, 55g boric acid and 40ml 0.5M EDTA. The agarose and 0.5xTBE were heated in a microwave to dissolve the agarose. Fifteen microlitres of SYBRsafe® DNA stain (Invitrogen, Life Technologies) was added to the agar solution after it was cooled slightly. The agar solution was poured into a gel tray and left to set with combs to demarcate the lanes.

Five microlitres of each of the PCR products were mixed with 5µl Cresol dye (0.25% Cresol red, 40% sucrose) and loaded onto each lane. The molecular size marker loaded was a 100bp DNA ladder (Biolab, London, UK). The gel was run for 30-45 minutes (depending on the expected size of the PCR product) at 150V (400mA) and imaged by the GeneGenius Bioimaging system (Syngene, Cambridge, UK). The annealing temperature that gave the strongest single band of the correct size was ascertained and used as the annealing temperature in all subsequent PCRs with that primer set. Subsequent PCRs were run as above, except that the programme used was specific to the target temperature required. To ensure that the PCR had worked before sending the samples for sequencing, several of the PCR products were sampled by agarose gel electrophoresis as above to check that a single product of the expected size was obtained.

2.3.10 Sequencing

Post PCR, samples were sequenced at the MRC Human Genetics Unit (Western General Hospital, Edinburgh), using the ABI 3730 Sanger sequencing machine (Applied Biosystems, Life Technologies). A single primer was used to synthesis new DNA fragments, but with specific fluorescent dyes attached to each nucleotide of the DNA strand of interest. A different dye was used for each of the four nucleotides: adenine, cytosine, guanine and thymine. The samples were run by capillary action through a gel matrix, so that there was a correlation between the length of the DNA and the time at the sensor, and this was used to determine the DNA sequence. The sequencing machine could detect the wavelength of each fluorescent dye using optical sensors.

2.3.11 Analysis of sequencing

Sequencing was analysed using Sequencher® 4.8-4.10 (Gene Codes Corporation, Ann Arbor, MI). The known sequence was downloaded from www.ensembl.org and the sequencing results were compared to this. The results were documented on a Microsoft® Excel spreadsheet.

2.3.12 Phylogenetic analysis of conservation

An analysis of conservation across species is helpful when considering a gene in which there may be an intronic mutation producing disease susceptibility, as areas that have been conserved across species are more likely to contain important regulatory non-coding areas.[213] Sequences of the gene of interest from different species were downloaded from www.ensembl.org in the FastA format and uploaded to PIPmaker (pipmaker.bx.psu.edu/pipmaker/), and percentage of identity plots (PIP) calculated. The multiple alignment programme of Vista was also used to complete a species alignment (pipeline.lbl.gov/cgi-bin/gateway2).

2.4 Cell culture

Unless otherwise stated, all plastic-ware was obtained from Greiner UK and cell culture reagents were obtained from Gibco, Life Technologies, UK. LS174T cells were used for gene expression analyses (section 2.5) whereas the SW480 cells were used for protein work (section 2.6)

2.4.1 Cell culture conditions

Both the SW480 and the LS174T colonic cell lines were obtained from the European Collection of Cell Cultures (Health Protection Agency, Salisbury, UK). The SW480 cells were grown in unvented flasks in L-15 Leibovitz medium supplemented with 10% foetal calf serum (FCS), 2mM glutamine, 100U/ml penicillin and 0.1mg/ml streptomycin. The LS174T cells were grown in vented flasks in minimum essential medium eagle (MEM) (Sigma-Aldrich) supplemented with 10% FCS, 1% non essential amino acids, 2mM glutamine, 100U/ml penicillin and 0.1mg/ml streptomycin. All cell lines were incubated at 37^{o}C in humidified air containing 5% CO_2.

2.4.2 Cell Passage

Cells were passaged by washing the cells several times with phosphate buffered saline (PBS) and incubating with 0.25% trypsin/EDTA for several minutes. The flask was then vigorously agitated to release the adherent cells. These cells were pooled in a falcon tube, spun down to a pellet and resuspended in fresh medium. For the LS174T cells, the cells were passed through a fine needle to ensure a single cell suspension. The cells were cut back in a ratio between 1:2 and 1:8 depending on when and how many flasks of cells were needed for forthcoming experiments.

2.5 Gene expression analyses

2.5.1 Transfection

Harvested LS174T cells of low passage number (<10) were counted using a haemocytometer and $1x10^6$ cells plated out into each cell of a 6 well plate, and medium added to give a total volume of 2ml. They were incubated to achieve 80-90% confluence at the start of the transfection. For each well to be transfected, 10µl lipofectamine™ 2000 (Invitrogen, Life Technologies) was diluted into 250µl Opti-MEM® (Invitrogen, Life Technologies) without serum and incubated for 5 minutes at room temperature. Then 2.5µg of the DNA to be transfected was added to 250µl Opti-MEM® and this mixture added to the lipofectamine™/Opti-MEM® that had already been made up. This was incubated at room temperature for 20 minutes to allow the DNA-lipofectamine™ complexes to form. After the above incubation, the mixture was added to wells already containing 1ml Opti-MEM®. The cells were incubated at 37°C for 5-6 hours and the complexes were removed and complete medium, as defined in 2.4.1, was added.

2.5.2 Time courses

Time courses were carried out with LS174T cells to examine the effects of different compounds on the cells. If a transfection was required (see above) the cells were allowed to recover for about 12 hours. Using cells in 6-well plates at 80-90% confluence, the medium was changed. Cells were stimulated with: TNFα (final concentration 50ng/ml), LPS (final concentration 1µg/ml), MDP (final concentration

1μg/ml), monensin (final concentration 1μM) and carbachol (final concentration 1mM) with a further well left unstimulated. The cells were incubated at 37°C for 8, 24 or 48 hours . To harvest, the cells were washed twice with 1ml PBS and then scraped in 1ml PBS and recovered into a 1.5ml tube. This was spun at 13,500rpm for 5 minutes, the supernatant removed and discarded, and the cell pellet stored at -80°C.

2.5.3 RNA extraction

RNA extraction was completed with the Qiagen RNeasy Mini kit (Qiagen, Crawley, UK), according to standard protocols. Briefly, cells were lysed with 350μl lysis buffer (RLT containing 10μl β-mercaptoethanol per ml) and mixed well. The lysate was pipetted onto a QIAshredder spin column in a 2ml collection column and centrifuged for 3 minutes at >8000g to homogenise the lysate. 350μl of 100% ethanol was added to the lysate and transferred to an RNeasy mini spin column placed in a 2ml collection tube. This was centrifuged for 15 sec at >8000g to allow the RNA to adsorb to the silica-gel membrane, and the liquid from the column (follow through) was discarded.

350μl Buffer RW1 was added to the column and centrifuged for 15s at >8000g, with the follow through being discarded. 80μl DNase (made up of 10μl DNase I stock solution with 70μl DNA digest buffer RDD) was added directly onto the column membrane and incubated at room temperature for 30-45 minutes to ensure removal of any remaining DNA.

Following this incubation a further wash was performed with 350μl Buffer RW1 and centrifuged for 15s at >10000rpm. 500μl Buffer RPE was added to the column and this was centrifuged for 2 min at >8000g to wash spin column membrane. The column was then removed into a new 1.5ml collection tube. 50μl RNase-free water was added directly onto the membrane and the spin column centrifuged for 1 min at >8000g to elute the RNA into the collection tube. The RNA was quantified on the Nanodrop and stored at -80°C.

2.5.4 cDNA synthesis

cDNA was made using the Invitrogen Superscript® III First-strand synthesis protocols and reagents (Invitrogen, Life Technologies). 1µg RNA was added to 1µl 10mM dNTP mix and 50ng random hexamers, and the total volume made up to 10µl with DEPC-treated water. This mixture was incubated at 65°C for 5 minutes to denature the RNA and primers, and then placed on ice for at least 1 minute. The cDNA synthesis mix was added to the mixture (2µl 10X RT Buffer, 4µl 25mM $MgCl_2$, 2µl 0.1M DTT, 1µl RNase OUT and 1µl Superscript® III reverse transcriptase and incubated at 25°C for 10 minutes to allow annealing. cDNA synthesis was promoted by incubating at 50°C for 50 minutes. The reaction was terminated by heating to 85°C for 5 minutes. The cDNA was stored at -80°C.

2.5.5 qPCR

Real-time or quantitative (q)PCR is a method of not only detecting, but also quantifying the amount of specific nucleotide sequences. A probe that preferentially binds to dsDNA and only fluoresces when bound to DNA was used in the PCR. The fluorescence intensity was measured at each PCR cycle allowing DNA concentrations to be quantified by comparing to a standard dilution of DNA.

2.5.5.1 SYBR® Green qPCR

Appropriate primers for each gene of interest were designed by Ms Kimberley Soo using the Primer3 programme (http://frodo.wi.mit.edu/primer3). An initial standard curve PCR was run to check that the primers were working and to calculate the cDNA optimal concentration for those primers. Serial cDNA dilutions were made:

1:5	1:10
1:50	1:100
1:500	1:1000
1:5000	1:10000
0	

RotorGene-specific 0.1ml strip tubes (Qiagen) were used to add 2µl of the relevant cDNA dilution, 10µl SYBR® Green EXPRESS (Invitrogen, Life Technologies), primer pairs at a working concentration of 1mM each and the mixture made up to 20µl with deionised water . Each sample was done in duplicate. The samples were loading into a 72-well rotor in the Rotor-Gene 6000 qPCR machine (Qiagen) and run in the conditions in Table 2-6.

Hold at 90°C for 2 minutes	
95°C for 15 seconds	}Repeated 40 times
60°C for 60 seconds	}

Table 2-6 qPCR conditions

The optimal concentration to run further qPCRs was calculated by choosing the concentration of cDNA which was within the exponential doubling phase of the qPCR. For each qPCR run, the R^2 was checked. R^2 is the percentage of the data which is consistent with the hypothesis that the given standards form a standard curve, and ideally should be around 0.99. The reaction efficiency is a mark of how efficient each cycle of amplification is (i.e. does the fluorescent signal double with each cycle), and if it is 100% efficient the value is 1.

Following the calculation of the optimal concentration for the qPCR, time course cDNAs at the appropriate optimal concentration were run with the same master mix as already listed. Each PCR also had 5 standards run as a control. For each set of cDNAs, the gene of interest was normalised to a glyceraldehyde-3-phosphate dehydrogenase (GAPDH) qPCR.

2.5.5.2 Taqman® qPCR

Previous work in the GI Unit by Ms Kimberley Soo had established that NOD2 qPCR was best achieved using the Taqman® system (Applied Biosystems, Life Technologies) rather than the SYBR® Green system. This involved the use of Taqman®'s inventoried, commercially available NOD2 and GAPDH assays. Serial dilutions of the cDNA were made to calculate the optimal concentration to run the qPCR:

Neat	1:5
1:10	1:50
1:100	1:500
1:1000	1: 5000
1:10000	1:50000

In addition, a Myc-tagged plasmid containing NOD2 (pCMV-Myc, BD Bioscience) was also used in serial dilutions for the standard curve:

1:1000000	1:5000000
1:10000000	1:50000000
1:100000000	1:500000000
1:1000000000	

RotorGene-specific 0.1ml strip tubes (Qiagen) were used to add 2µl of the relevant cDNA dilution, 5µl Taqman® Gene Expression master mix (Applied Biosystems, Life Technologies), 0.5µl primer mix (premixed assay of the relevant primers with the FAM reporter dye label, Applied Biosystems, Life Technologies) and deionised water to make up to 10µl. Each sample was done in duplicate. The samples were loading into a 72-well rotor in the Rotor-Gene 600 qPCR machine (Qiagen) and run on the programme shown in Table 2-7.

Hold at 50°C for 2 minutes
Hold at 95°C for 10 minutes
95°C for 15 seconds } Repeat 40 times 60°C for 60 seconds }

Table 2-7 Taqman® qPCR conditions

Following the calculation of the optimal concentration for the qPCR, time course cDNAs at the appropriate optimal concentration were run with the same master mix as already listed. Each PCR also had 5 standards of NOD2 plasmid run as a control. For each set of cDNAs, the gene of interest was normalised to a GAPDH Taqman® qPCR.

2.6 Protein work

2.6.1 Transfection

SW480 cell transfections for western blotting and co-immunoprecipitation (Co-IP) were completed in T25 flasks incubated to achieve 80-90% confluence at the start of the transfection. For each well to be transfected, 20µl lipofectamine™ 2000 (Invitrogen, Life Technologies) was diluted into 500µl Opti-MEM® (Invitrogen, Life Technologies) without serum and incubated for 5 minutes at room temperature. Then 5µg of the DNA to be transfected was added to 500µl Opti-MEM® and this mixture added to the lipofectamine™/Opti-MEM® that had already been made up. This was incubated at room temperature for 20 minutes to allow the DNA-lipofectamine™ complexes to form. After the above incubation, the mixture was added to wells already containing 1.5ml Opti-MEM®. The cells were incubated at 37°C for 5-6 hours and the complexes were removed and complete medium, as defined in 2.4.1, was added. After further incubation for at least 12 hours, the cells were harvested by being washed twice with 1ml PBS, scraped in 1ml fresh PBS and recovered into a 1.5ml tube. This was spun at 13,500rpm for 5 minutes, the supernatant removed and discarded, and the cell pellet stored at -80°C.

2.6.2 Cell lysate preparation

Protein lysate for western blotting was produced by lysing the cell pellet with 500µl NP40 lysis buffer (master mix containing 16ml water, 1ml 2M NaCl, 1ml 10% NP-40 (Calbiochem®, Merck, UK), 1ml 1M Hepes Buffer, 400µl 0.5M EDTA and 800µl complete protein inhibitor mix (Roche, UK)). The solution was incubated on ice for 20 minutes and spun at 13,000g at 4°C for 15 minutes. The supernatant was transferred to a fresh tube ready for western blotting or co-immunoprecipitation, and stored at -80°C if not immediately needed.

2.6.3 Protein level quantification

Protein lysate protein levels were quantified using the Qubit™ protein assay kit (Invitrogen, Life Technologies).

2.6.4 Gel electrophoresis and western blotting

Gel electophoresis used the NuPAGE® gel electrophoresis system (Invitrogen, Life Technologies). For western blotting with protein lysates, 20µg protein lysate was added to 5µl 4x NuPAGE® LDS sample buffer, 2µl 10x NuPAGE® reducing agent along with sufficient deionised water to give a final volume of 20µl. This sample was heated to 70°C for 10 minutes to denature the protein.

1x NuPAGE® MOPS SDS running buffer was made up with the 20x solution provided and deionised water. A NuPAGE® 4-12% Bis-Tris gel was mounted in the XCell Surelock™ Minicell system and the running buffer added into both chambers. 500µl NuPAGE® antioxidant was added to the inner buffer chamber. The denatured protein lysate samples were loaded onto the gel, with protein markers (a mixture of 10µl Magic Marker and 5µl See Blue, both from Invitrogen, Life Technologies) in one lane. The samples were run at 200V for 50 minutes.

Following electrophoresis, western blotting was immediately started. 1x NuPAGE® Transfer buffer was made up with the 20x solution provided, deionised water, 10% methanol (20% if transferring 2 gels) and 0.1% NuPAGE® antioxidant. The transfer pads and the filter paper were soaked in this buffer. The PVDF membrane was presoaked in methanol for 30 seconds before being soaked in the transfer buffer. The membrane was placed on the post electrophoresis gel and this was sandwiched between the filter paper. This was put into the blot module surrounded by the transfer pads, according to Invitrogen's protocols, and the transfer run for 1 hour at 30V. Following transfer the membrane was blocked in a milk solution (5% w/v Marvel (Premier Foods, UK) in PBS with 0.1% v/v Tween 20) for at least 1 hour at room temperature.

2.6.5 Co-immunoprecipitation

For each co-immunoprecipitation (CoIP) reaction, identical amounts of protein lysate were used. In order to achieve this, 400µl of the lowest protein level sample was used as the quantity of protein to be used in each reaction. Appropriately calculated amounts of the other samples were used and each diluted with deionised water to achieve identical final concentrations and volumes.

Protein G agarose beads (Roche, UK) were washed twice with PBS and centrifuged after each wash at 13000g for 30 seconds. An equal volume of PBS was added to the agarose beads, and 20-30µl of the bead slurry pipetted out for each CoIP. An appropriate amount of antibody (usually 2µg) and 400µl protein lysate was added to the bead slurry. The mixture was incubated overnight on an orbital shaker at 4°C.

The agarose beads were collected by short pulses of centrifugation and the beads washed twice with PBS. PBS was added and incubated at 4°C for 10 minutes on an orbital shaker, and the beads centrifuged for 30 seconds. The agarose beads were resuspended in 20µl 2x sample buffer, made up of 250µl 4x NuPAGE® LDS sample buffer, 200µl PBS and 50µl β mercaptoethanol. The agarose beads were boiled for 3 minutes to dissociate the immunocomplexes from the beads and denature the protein. After centrifugation at 13000g for 2 minutes the supernatant was collected and used to proceed directly to loading on the gel for gel electrophoresis and western blotting as described in section 2.6.4. Protein lysate samples were also loaded onto the gel as described in section 2.6.4 to demonstrate that the initial transfection had been successful.

2.6.6 Protein probing

The transfer membrane from the western blotting was incubated overnight with the appropriate primary antibody at a concentration of 1µg/ml in 5% w/v Marvel in PBS with 0.1% v/v Tween 20 and constant agitation. The membrane was washed 3 times for 10 minutes each in PBS/0.1% v/v Tween for 10 minutes and the appropriate secondary antibody applied in a concentration of 1:1000 in 5% w/v Marvel in PBS with 0.1% v/v Tween 20 (eg goat anti-rabbit IgG-HRP (Santa Cruz Biotech, Santa Cruz, CA) if a rabbit primary had been used) and incubated for 2 hours at room temperature before washing. Finally the membrane was incubated in Immobilon™ Western Chemiluminescent HRP substrate (Millipore, Billerica, MA) for 5 minutes. X-ray film (Z370371, Sigma) was placed over the membrane for 10 seconds - 2 minutes in a darkroom and the film passed through an X-ray developer to view the luminescence of the membrane, seen as bands. The markers were used to check that the bands seen were the correct size for the protein of interest.

2.7 Expression studies on human intestinal tissue

2.7.1 Recruitment of patients

Patients attending for routine colonoscopy were approached to give consent for their participation in the study using a standard letter sent to them 1-2 weeks before their appointment (Appendices 10.7 and 10.8). Patients with IBD were defined according to the standard criteria given in section 2.2.5. Controls were recruited from the patient population attending for surveillance for bowel cancer or colonic polyps.

2.7.2 Ethics

The study had the full approval of the Lothian LREC (2001/4/72).

2.7.3 Biopsy protocol

Two biopsies were taken from each of: terminal ileum (if reached), ascending colon, transverse colon, descending colon, sigmoid colon and rectum. Biopsies were collected and fixed in formaldehyde, mounted in paraffin blocks and sectioned onto Superfrost® Plus slides (Menzel Glaser, Germany). The fixing, mounting and sectioning were done either in the Pathology Department, Western General Hospital, or at the Breakthrough Research Unit, Western General Hospital.

2.7.4 Immunohistochemistry protocol

Sections were deparaffinised in xylene (2 incubations of 5 minutes each), rehydrated in graded ethanol washes (2 x 5 minute washes in 100% ethanol then 2 x 5 minute washes in 70% ethanol) and finally washed in running water for 5 minutes. Heat Induced Epitope retrieval (HIER) was performed in 500ml citrate buffer (10mM citric acid, pH adjusted to 6.0 with 1M NaOH, then 0.25ml Tween-20 added) in a pressure cooker for 5 minutes. After a further wash for 5 minutes the slides were loaded into a Sequenza® (Thermo Scientific, UK) and washed in PBS for 5 minutes. The Envision™ system (Dako, Ely, UK) was used according to their protocols. Briefly, the dual endogenous enzyme block was applied for 5 minutes to block endogenous peroxidase activity. After a 5 minute wash with PBS, 100μl of the appropriately diluted primary antibody was applied and left to incubate for at least 30 minutes. After a further wash in PBS for 5 minutes the slides were incubated for 30 minutes in 3 drops of labelled polymer HRP (horseradish peroxidase) – the secondary antibody. The slides had a further wash in PBS for 5 minutes and were unloaded from the Sequenza®. 3,3'-diaminobenzidine (DAB) was made up using 1ml of substrate buffer and 20µl DAB chromogen; 3 x 2 minute treatments were applied of the DAB with a brief wash in running water in between. The slides were washed in running water, counterstained with filtered Mayer's haematoxylin for 5 minutes and dipped 3 times very briefly into 1% HCl in 70% ethanol to remove excess haematoxylin. The slides were put in 0.2M lithium carbonate to 'blue' the slides, following a quick wash they were dehydrated in graded ethanol (2 x 5 minute washes in 70% ethanol then 2 x 5 minute washes in 100% ethanol) and mounted with pertex mounting medium. Slides were visualized and imaged on an Olympus microscope.

2.8 Site Directed Mutagenesis

2.8.1 Initial PCR

The Quikchange® site-directed mutagenesis kit (Stratagene®, La Jolla, CA) was used to attempt to introduce the 908 and 1007fs mutations into the NOD2 gene. Wild-type NOD2 had previously been cloned into a pCMV-Myc vector (BD Bioscience, gift of Dr Elaine Nimmo), and appropriate primers for the site-directed mutagenesis NOD2 mutations had already been designed by Dr Elaine Nimmo. A PCR was run with 5µl 10x reaction buffer, 1µl dNTP mix, 2-10µl pCMV-Myc vector with NOD2wt, 125ng of each of the primers, 1µl *PfuTurbo* DNA polymerase (2.5U/µl) and made up to 50µl with ultra-pure water. A control PCR was also set up with 5µl 10x reaction buffer, 1µl dNTP mix, 2µl (10ng) pWhitescript™ control plasmid, 1.25µl of each of the control primers, 1µl *PfuTurbo* DNA polymerase (2.5U/µl) and sufficient double distilled water to give a reaction volume of 50µl. All the standard reagents were from the Quikchange® kit. A PCR was performed with NOD2wt DNA according to the conditions in Table 2-8. The control plasmid PCR was run according to the conditions in Table 2-9.

95°C for 30 seconds	
95°C for 30 seconds	} Repeated 16 times
55°C for 1 minute	}
68°C for 4 minutes (1minute/Kbp plasmid length)	}

Table 2-8 Site-directed mutagenesis PCR for NOD2wt DNA

95°C for 30 seconds	
95°C for 30 seconds	} Repeated 18 times
55°C for 1 minute	}
68°C for 5 minutes (1minute/Kbp plasmid length)	}

Table 2-9 Site-directed mutagenesis PCR from control plasmid

Following PCR both reactions were placed on ice for 2 minutes to cool the reaction to <37°C. 1µl *Dpn* I restriction enzyme (10U/µl) was added to each reaction, mixed well and incubated at 37°C for 1 hour. This enzyme is specific for methylated and hemi-methylated DNA and thus digests the parental DNA template whilst leaving the mutation-containing synthesised DNA intact, which contains staggered nicks due to incorporation of the primers.

2.8.2 Transformation of XL-1 Supercompetent cells

The XL1-blue *E.Coli* cells were gently thawed on ice. For each control and sample transformation, 50µl of the cells were aliquoted into a prechilled 14ml polypropylene round-bottom tube (BD Biosciences), and 1µl of the appropriate post-PCR *Dpn* I treated DNA added to the tube. As a control for this step, the transformation efficiency of the XL-1 cells was checked by adding 1µl of the pUC18 control plasmid (concentration 0.1ng/µl) to a further 50µl aliquot of the supercompetent cells. The transformation reactions were gently mixed and incubated on ice for 30 minutes. The transformation reactions were heat pulsed for 45 seconds at 42°C and then put on ice for 2 minutes. 500µl NZY$^+$ broth, having been preheated to 42°C, was added to each reaction and the mixture incubated at 37°C for 1 hour with shaking at 225-250 rpm.

The NZY^+ broth was made as follows. 10g NZ amine (casein hydrolysate), 5g yeast extract, 5g NaCl were mixed and the solution made up to 1 litre with deionised water. After adjusting the pH to 7.5 with NaOH and autoclaving the solution, 12.5ml 1M $MgCl_2$, 12.5ml 1M $MgSO_4$ and 20ml 20% (w/v) glucose was added.

Following incubation, 250µl aliquots of the cells were individually spread on Lysogeny broth (LB)-ampicillin (0.1mg/ml ampicillin) agar plates containing 80µg/ml 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) and 20mM isopropyl-1-thio-β-D-galactopyranoside (IPTG). For each of the sample mutagenesis reactions, 2 aliquots of the cells were individually plated out so that the whole of the 500µl transformation reaction was used. Only 5µl of the pUC18 control mixture was plated out using a further 200µl NZY^+ broth. The plates were incubated at 37°C overnight.

2.8.3 Mutagenesis efficiency

The pWhitescript™ control plasmid contains a stop codon (TAA) at the position where a CAA codon (coding for glutamine) would normally appear in the β galactosidase gene of the plasmid. XL1-blue supercompetent cells transformed with this control plasmid appear as white colonies on LB-ampicillin plates containing IPTG and X-gal, as the β galactosidase activity has been removed. However the point mutation introduced by the control primers revert the stop codon back to CAA and thus colonies can be screened for the β galactosidase (β gal^+, blue) phenotype. The expected colony number from the transformation of the pWhitescript™ control mutagenesis reaction was 50-800 colonies. More than 80% of the mutagenesis control colonies needed to contain the mutation for the site-directed mutagenesis control PCR to have been successful, and this was gauged by the number of blue colonies on the mutagenesis control agar plates containing IPTG and X-gal.

2.8.4 Transformation efficiency

The transformation efficiency was calculated using the pUC18 control plasmid which was expected to give >250 colonies, with >98% having the blue phenotype, meaning that the transformation efficiency was $>10^8$cfu/µg.

2.8.5 NOD2 mutagenesis analysis

Individual colonies were streaked out onto an agar plate (with 100µg/ml ampicillin) using a pipette tip, and the pipette tip was dipped afterwards into the pre-prepared PCR mix to provide the DNA template for a PCR reaction to amplify the area of the expected mutagenesis, which was sequenced to check for successful introduction of the mutation. Where the mutation appeared to have been introduced, the entire NOD2 gene was sequenced (using the methods detailed in sections 2.3.10 and 2.3.11), to check that no other mutations had been introduced in the process.

2.8.6 Plasmid DNA generation

Colonies containing the introduced mutation were inoculated into 5ml LB culture medium (made by suspending 10g tryptone, 5g yeast extract and 10g NaCl in a total volume of 1 litre made up with water) with 0.1mg/ml ampicillin and incubated for about 7 hours at 37°C with gentle agitation. 10µl of this solution was inoculated into 3ml LB medium with 0.1mg/ml ampicillin and the solution incubated overnight at 37°C.

2.8.7 Plasmid DNA purification

Plasmid purification was performed according to the Qiagen Plasmid Purification Midi kit protocols (Qiagen). After overnight culture, 25ml bacterial cell suspension was centrifuged at 6,000g for 15 minutes at 4°C, and the supernatant discarded. The resulting pellet was resuspended in 4ml buffer P1 (resuspension buffer), and 4ml buffer P2 (lysis buffer) added and mixed well. 4ml buffer P3 (neutralisation buffer) was added and mixed well, and the mixture incubated on ice for 15 minutes. This solution was centrifuged at >20,000g for 30 minutes at 4°C and the supernatant centrifuged again at >20,000g for 15 minutes at 4°C. A Qiagen–tip 100 was equilibrated by adding 4ml buffer QBT (equilibration buffer) and allowing the column to empty by gravity flow. The centrifuged supernatant was applied to the

Qiagen-tip and allowed to enter the resin by gravity flow. Following this the Qiagen-tip was washed twice with 10ml Buffer QC (wash buffer), which was also allowed to move through the tip by gravity flow. The DNA was eluted using 5ml Buffer QF (elution buffer) into a 15ml tube. The eluted DNA was precipitated by adding 3.5ml room temperature isopropanol and mixing, with subsequent centrifugation at >15,000g for 30 minutes at 4°C. The supernatant was decanted, and the resulting DNA pellet washed with 2ml room temperature 70% ethanol and centrifuged at >15000g for 10 minutes, and the supernatant removed again. The pellet was air dried for 10 minutes and the DNA redissolved in an appropriate volume of 1x TE.

2.8.8 Plasmid DNA glycerol stock production

100µl *E.Coli* K-12 cells (known competency 1×10^8 cells/ml) and approximately 100ng plasmid DNA were mixed in a 1.5ml tube and incubated on ice for 30 minutes. This was heated to 42°C for 2 minutes and then incubated on ice for 1 minute. 500µl LB culture medium with 0.1mg/ml ampicillin was added and incubated at 37°C for 1 hour. 100µl of this solution was spread onto an Agar plate (containing 100µg/ml ampicillin) and incubated overnight at 37°C. Single colonies were picked from each plate, put into 2ml LB culture medium with 100µg/ml ampicillin, and incubated at 37°C for 8 hours. The samples were spun for 2 minutes at >20000g and the supernatant removed. 700µl LB culture medium with 25% glycerol was added to the pellet and the stocks stored at -20°C.

2.9 Statistics

2.9.1 Individual SNPs

Individual SNPs were analysed through Haploview Version 4.2 or with the JMP 8.0 (SAS) statistical package using the chi-squared or Fisher's exact test, as appropriate. A significant p-value was considered to be <0.05, but where many SNPs were genotyped, a simple Bonferroni correction[214] was made for multiple testing.

2.9.2 Gene-wide haplotype tagging methodology

Multiple tagging SNPs were selected for genes based on HapMap data loaded onto Haploview and analysed according to the tagger algorithm embedded in Haploview version 4.2.[215]

2.9.3 Haplotype analysis

Haplotypic analyses were completed in Haploview version 4.2.

2.9.4 Power calculations

Power calculations for genetic studies were completed using Quanto[216] (http://hydra.usc.edu/GxE/).

2.9.5 Kaplan-Meier survival analyses

Kaplan-Meier analysis is a method of estimating survival function from patient data, and when the patient population is large enough, approximates the true survival function for that population.[217] In the context of this thesis it was used to estimate the proportion of patients who had had a progression of their disease or had required an operation at any particular time point in disease course, bearing in mind 'right censoring' (patients not followed up long enough to have developed the complication). Analyses of the Dundee and Edinburgh cohorts were performed using Microsoft Excel™ and GraphPad Prism version 4.00 (GraphPad Software, San Diego California USA), with the help of Dr Nicholas Lewin-Koh.

2.9.6 Tests of correlation

Tests of correlation were completed using GraphPad Prism version 4.00 for Windows (GraphPad Software, San Diego California USA).

2.9.7 Chi-squared test

Comparisons between groups were made using the chi-squared test, or Fisher's exact test where appropriate, using using GraphPad Prism version 4.00 for Windows (GraphPad Software, San Diego California USA).

2.9.8 Odds ratios

All odds ratios are presented with 95% confidence intervals.

2.9.9 Receiver operating characteristic curve

The receiver operating characteristic (ROC) curve is a method for displaying the relationship between the sensitivity and specificity of a continuous variable to predict another binary variable. ROC curve analyses were completed using both GraphPad Prism version 4.00 for Windows (GraphPad Software, San Diego California USA) and JMP version 8.0.2 (SAS, Cary, NC, USA).

2.9.10 Multivariate analyses

Multivariate analyses on survival function were completed at Genentech, San Francisco by Nicholas Lewin-Koh using a Cox Proportional Hazards model designed in the R programme (version 2.10.1).

The multivariate analyses on the severity data were completed using the statistical packages embedded in JMP version 8.0.2 (SAS, Cary, NC, USA).

Chapter 3 Analysis of disease progression and need for surgery in a Scottish Crohn's Disease cohort

Summary

Aim: To define the CD phenotype in Scottish patients according to the Montreal classification by examining the risk of disease progression and risk of surgery, including the need for multiple surgical resections.

Methods: A total of 1155 patients recruited from Edinburgh and Dundee were examined. Time to disease progression (development of stricturing and/or penetrating disease) and time to first resection were examined by Kaplan-Meier analyses. Univariate analyses were completed on these factors with respect to disease location, age at diagnosis, smoking at diagnosis, perianal disease, and, for time to first resection, decade at diagnosis. Multivariate analyses of time to disease progression and time to first resection were also completed using a Cox Proportional Hazards model. Kaplan-Meier analysis was used to examine the need for multiple resections with respect to disease location. Patients who had developed stricturing disease as their first evidence of disease progression were examined for their subsequent risk of developing penetrating disease, and those with penetrating disease as their first evidence of disease progression were examined for their subsequent risk of stricturing disease.

Results: Median time to disease progression was 14.5 years (95% CI 10.5-18.6). On univariate analyses L2 disease had a longer median time to disease progression than other disease locations. Age at diagnosis, smoking at diagnosis and presence of perianal disease were not important on univariate analyses. This was confirmed on a multivariate analysis where the only significant factor was disease location, with L1 disease conferring a HR of 4.7 (95% CI 3.6-6.1) and L3 disease HR 2.8 (95% CI 2.1-3.8) compared with L2 disease. The median time to first resection was 8.9 years (95% CI 7.5-10.0). On univariate analyses L2 disease and those diagnosed in a later decade had a longer median time to first resection. Age at diagnosis, smoking at diagnosis and presence of perianal disease were not important on univariate analyses. On a multivariate analysis (which excluded decade at diagnosis) the only significant factor for risk of first resection was disease location with L1 disease conferring a HR of 5.2 (95% CI 4.1-6.5) and L3 disease HR 2.6 (95% CI 2.1-3.3) compared with L2 disease. On Kaplan-Meier analysis of time to subsequent resections, there was a

statistically significant trend to increasing time between subsequent operations ($p<0.0001$). This appeared to be primarily driven by patients with L2 disease. Risk of stricturing disease in those with penetrating disease was similar to the penetrating disease in those with stricturing disease (p 0.843 on log rank test)

Conclusions: This study confirms the findings of previous studies that disease location is an important factor in determining risk of disease progression and risk of resection. It also demonstrates that disease location is important in determining the need for multiple resections. The data on disease behaviour would suggest that the stricturing and penetrating categories of disease behaviour should be considered as separate variables.

3.1 Introduction

Crohn's disease (CD) is an incurable inflammatory disease characterised by the tendency to progress to stricturing and/or penetrating disease. Stricturing and penetrating disease can be major causes of morbidity for patients, with the need for intestinal resections as well as other surgical procedures such as stricturoplasties and incision and drainage of abscesses.

The Montreal classification[20], a modification of the Vienna classification[17] is a method of classifying CD into meaningful categories, useful in research studies to provide uniformity of definitions across different clinicians and research groups. A summary of the Montreal classification is shown in Table 3-1. In particular, disease behaviour is commonly defined at a particular time point, usually 5 years after diagnosis, although in the analysis presented in this chapter a defined time point was not used because time to change in behaviour was examined.

Age at diagnosis	
A1	<17 years old
A2	17-40 years old
A3	>40 years old
Disease location (max extent before 1st resection)	
L1	Ileal +/- caecal overspill
L2	Colonic
L3	Ileocolonic
L4	Upper GI (+L1-L3 with concominant UGI disease)
Disease behaviour	
B1	Inflammatory
B2	Stricturing
B3	Internally penetrating
+/- perianal modifier 'p'	

Table 3-1 Montreal classification of Crohn's disease[20]

The problem with attempting to classify any disease as heterogeneous as CD is that a balance needs to be found between adequately describing disease variables and making the classification too complex to apply consistently. Therefore, by necessity, the Montreal classification can only be considered to give an outline of disease phenotype.

One major area for debate is the behaviour variable, where penetrating disease (B3) overrides stricturing disease (B2). This is based on the theory that increased intraluminal pressure (such as can occur in stricturing disease) helps to promote fistulae formation, reinforced by a small study of 42 consecutive CD intestinal resection samples in an Austrian hospital.[218] Of these patients, 27 had internal fistulae, 26 of which had co-existent stricturing in the surgical specimen. A bigger study of 236 resection specimens found 60 specimens with fistula formation of which only 4 did not have stricturing within the surgical specimen.[219] Neither of these studies gives any details of the age at diagnosis of the patients in the study, although it seems likely that the majority of patients had adult onset disease. Thus it is generally been assumed that patients with penetrating disease will have co-existent stricturing disease. In contrast, however, a study of 55 paediatric CD resection specimens demonstrated that only 16 of the 28 specimens with fistulae formation also had evidence of stenosis in the surgical specimen.[220] Whether this represents a difference between the paediatric and adult CD phenotype is unclear, but it certainly merits further investigation. As discussed in the introduction, the recently published Paris classification[21], a paediatric modification of the Montreal classification, has introduced a separate B2B3 disease behaviour category in recognition of the fact that in paediatric patients, stricturing and penetrating behaviours may be not be linked.

Other factors not accounted for in the Montreal classification and purposely omitted from the preceding Vienna classification in the interests of simplicity[17] include extraintestinal manifestations and family history, as well as a code for the oral disease location. Furthermore there are additional factors that may be important in defining severe disease: nutritional status, drug response, socio-economic impact of disease and the need for multiple resections that are not within the Montreal classification. Debate continues as to what classifies ‘severe’ disease, with no overriding consensus, but this will be addressed in more detail in the next chapter.

The aim of this chapter was to define the CD phenotype in Scottish patients according to the Montreal classification by examining the risk of disease progression and risk of surgery, alongside the risk of multiple resections.

3.2 Patient recruitment

Patients were recruited and phenotyped as detailed in Chapter 2. The majority of patients were recruited some time after their disease had been initially diagnosed (median time from diagnosis to recruitment 9.4 years, interquartile range 3.2-18.8 years).

3.3 Definitions

3.3.1 Stricturing disease

The definition of stricturing disease was according to the Vienna[17] and Montreal[20] classifications: 'The occurrence of constant luminal narrowing demonstrated by radiologic, endoscopic, or surgical examination combined with prestenotic dilation and/or obstructive symptoms'.

3.3.2 Penetrating disease

The definition of penetrating disease was according to the Montreal classification: 'The occurrence of intra-abdominal fistulas, inflammatory masses and/or abscesses at any time in the course of disease'.[20] Perianal fistulating disease was considered separately.

3.3.3 Disease location

As per the Montreal classification, disease location was defined as the maximum extent prior to the first resection. For patients who had not had a resection, it was defined as the maximum extent during follow up. The maximum extent was defined as the sum of all the areas affected at any point prior to the first resection, even if different areas had been affected at different times.

3.3.4 Time to last follow up

Time to last follow up was the time from diagnosis of CD to the last investigation (for example radiological or endoscopic examination) rather than the time to the last clinic appointment.

3.3.5 Time to disease progression

Time to disease progression was defined as the time from diagnosis to the first stricturing or internally penetrating complication.

3.3.6 Time to stricturing and penetrating disease

Patients developing stricturing disease as their first evidence of disease progression could subsequently develop penetrating disease; patients developing primarily penetrating disease were assumed to also have stricturing disease, as discussed in section 1.1.

3.3.7 Time to first surgical resection

Time to first surgical resection was defined as the time from diagnosis to the first intestinal resection, no matter how much intestine was resected. It excluded simple appendicectomies (even if CD was found in the resected specimen), defunctioning procedures and stricturoplasties.

3.4 Patient demographics

Basic patient demographics and comparisons (using chi-squared test) between the cohorts are shown in Table 3-2. Most patients (68%) were recruited in Edinburgh. There was a significant difference in the proportions of males in the two cohorts, (44.7% in the Dundee cohort vs. 37.3%, $p=0.02$). The median age at diagnosis was greater in the Dundee cohort, with more patients diagnosed at >40 years (35.7% vs. 26.0%, $p=0.0007$), and fewer patients diagnosed at <17 years old (A1, 7.4% vs. 11.7%, $p=0.0246$). Both cohorts had similar durations of follow-up. There was no significant difference in the proportions of current and non–smokers at diagnosis, but a higher proportion of the Dundee cohort were ex-smokers (17.2% vs. 11.2%, $p=0.0049$). The proportion of patients with L1 disease was lower in the Dundee cohort (19.6% vs. 30.6%, $p<0.0001$); the opposite was the case for L3 disease

(31.3% vs. 19.9%, p<0.0001). Although similar proportions of patients in both cohorts had inflammatory disease at 5 years, more patients had stricturing disease and fewer had internally penetrating disease in the Dundee cohort (stricturing disease 18.1% vs. 12.3%, p=0.0262); internally penetrating disease 10.8% vs. 16.4%, p=0.0345). The large percentage of 'not knowns' for disease behaviour was either due to records having been destroyed or because patients had been managed elsewhere and their full investigation profile was not available at 5 years.

Variable	Dundee cohort	Edinburgh cohort	Combined	Chi sq p-value
Total number of patients (% of total)	367 (32%)	788 (68%)	1155	NA
Sex (% male)	164 (44.7%)	294 (37.3%)	458 (39.7%)	0.02
Median age at diagnosis/years	30.8 (IQR 22.9-46.1)	27.8 (IQR 20.0-40.7)	28.6 (IQR 21.2-42.9)	
Age group at diagnosis (% of total)				
A1	27 (7.4%)	92 (11.7%)	119 (10.3%)	0.0246
A2	209 (56.9%)	491 (62.3%)	700 (60.6%)	0.0825
A3	131 (35.7%)	205 (26.0%)	336 (29.1%)	0.0007
Median duration of follow up/years	10.1 (IQR 4.1-19.7)	10.3 (IQR 4.1-19.8)	10.3 (IQR 4.1-19.7)	
Smoking at diagnosis (% of total)				
Current	150 (40.9%)	317 (40.2%)	467 (40.4%)	0.2466
Ex	63 (17.2%)	88 (11.2%)	151 (13.1%)	0.0049
Non smoker	152 (41.4%)	355 (45.0%)	507 (43.9%)	0.8356
Not known	2 (0.5%)	28 (3.5%)	30 (2.6%)	
Disease location (% of total)				
L1 Ileal	72 (19.6%)	241 (30.6%)	313 (27.1%)	<0.0001
L2 Colonic	118 (32.2%)	279 (35.4%)	397 (34.4%)	0.2784
L3 Ileocolonic	115 (31.3%)	157 (19.9%)	272 (23.5%)	<0.0001
L4 Upper GI	4 (1.1%)	24 (3.0%)	28 (2.4%)	0.0442
L4+another location	51 (13.9%)	69 (8.8%)	120 (10.4%)	0.0077
Not known	7 (1.9%)	18 (2.3%)	25 (2.2%)	
Disease behaviour at 5 years (259 Dundee patients and 560 Edinburgh patients) (% of total)				
B1 Inflammatory	165 (63.7%)	356 (63.6%)	521 (63.6%)	0.9702
B2 Stricturing	47 (18.1%)	69 (12.3%)	116 (14.2%)	0.0262
B3 Penetrating	28 (10.8%)	92 (16.4%)	120 (14.7%)	0.0345
Not known	19 (7.3%)	43 (7.7%)	62 (7.6%)	

Table 3-2 Dundee and Edinburgh basic demographics

Variable	Males	Female	Chi sq p-value
A1	57 (12.5%)	61 (8.8%)	0.047
A2	269 (58.7%)	434 (62.2%)	0.229
A3	132 (28.8%)	202 (29.0%)	0.953
L1	107 (23.3%)	206 (29.6%)	0.022
L2	168 (36.7%)	229 (32.9%)	0.184
L3	107 (23.4%)	165 (23.7%)	0.944
L4	16 (3.5%)	12 (1.7%)	0.077
L4+ another location	50 (10.9%)	70 (10.0%)	0.694
Unknown	10 (2.2%)	15 (2.1%)	

Table 3-3 Age group at diagnosis and Montreal locations according to sex

The analysis of age group at diagnosis and disease location according to sex is shown in Table 3-3. The only statistically significant differences between the sexes were with age at diagnosis and L1 disease.

Variable	Non	Ex	Current	Chi Sq p-value
L1	119 (23.4%)	39 (25.8%)	144 (30.9%)	0.032
L2	189 (37.3%)	68 (45.1%)	128 (27.5%)	<0.0001
L3	114 (22.4%)	26 (17.2%)	128 (27.5%)	0.023
L4	16 (3.2%)	5 (3.3%)	6 (1.3%)	0.119
L4+another location	60 (11.8%)	12 (7.9%)	46 (9.8%)	0.332
Unknown	9 (1.8%)	1 (0.7%)	14 (3.0%)	

Table 3-4 Analysis of disease location with respect to smoking status

The analysis of disease location with respect to smoking status at diagnosis (Table 3-4) demonstrated that L2 disease was more common in non- and ex-smokers than with current smokers, with the opposite being true for L1 and L3 disease.

3.5 Disease progression data

3.5.1 Time to disease progression

Including patients who had stricturing or penetrating disease at diagnosis, the median time to disease progression for the cohort was 14.5 years (95% confidence interval (CI) 10.5-18.6 years). The Kaplan-Meier graph of progression-free proportion is shown in Figure 3-1. At one year after diagnosis, most of the patients with stricturing or penetrating disease had already had this phenotype at diagnosis: 18% of patients had stricturing or penetrating disease at diagnosis, whereas at one year follow-up, 20.5% of patients had stricturing or internally penetrating disease. By 5 years this percentage was 32.6% and then increased to 45% at 10 years and 54.5% at 20 years, as shown in Table 3-5.

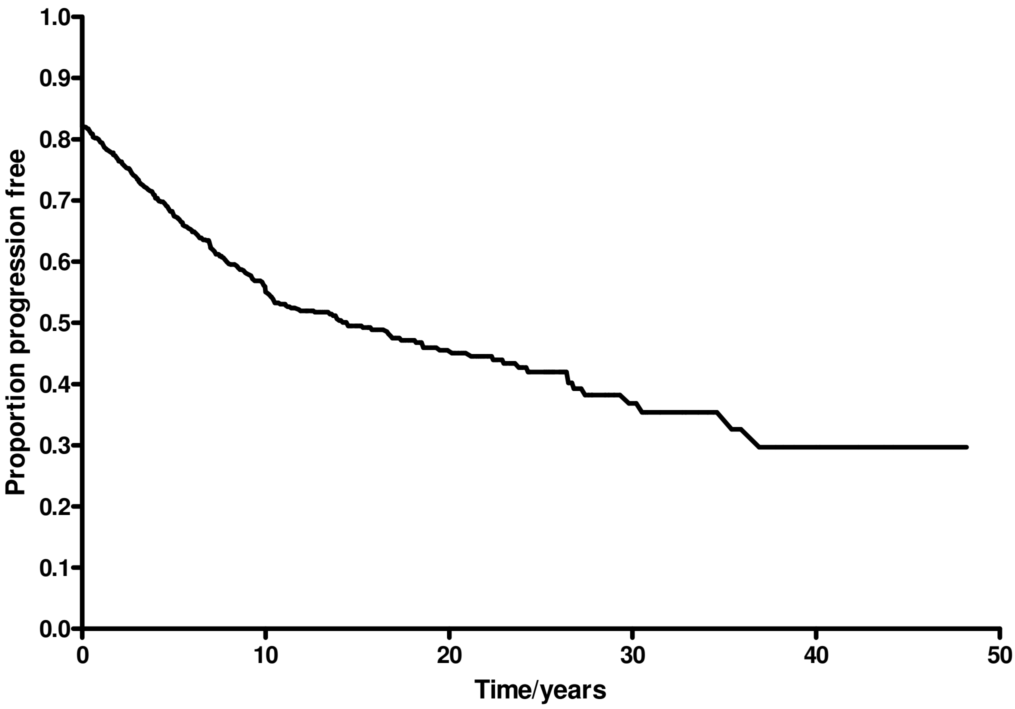


Figure 3-1 Kaplan-Meier of development of stricturing or penetrating disease

Time/years	% with stricturing or penetrating disease
0	18.0
1	20.5
3	26.5
5	32.6
10	45.0
20	54.5

Table 3-5 Percentages of patients with development of stricturing or penetrating disease, including those with disease progression at diagnosis

This analysis was repeated having excluded patients who had stricturing or penetrating disease at diagnosis; the Kaplan-Meier curve is shown in Figure 3-2. The median time to disease progression was 26.5 years. At one year after diagnosis, of those still under follow up, only 3.1% of patients with inflammatory disease at diagnosis had developed stricturing or penetrating disease. By 5 years, 17.8% of patients still under follow up had developed stricturing or penetrating disease, as shown in Table 3-6.

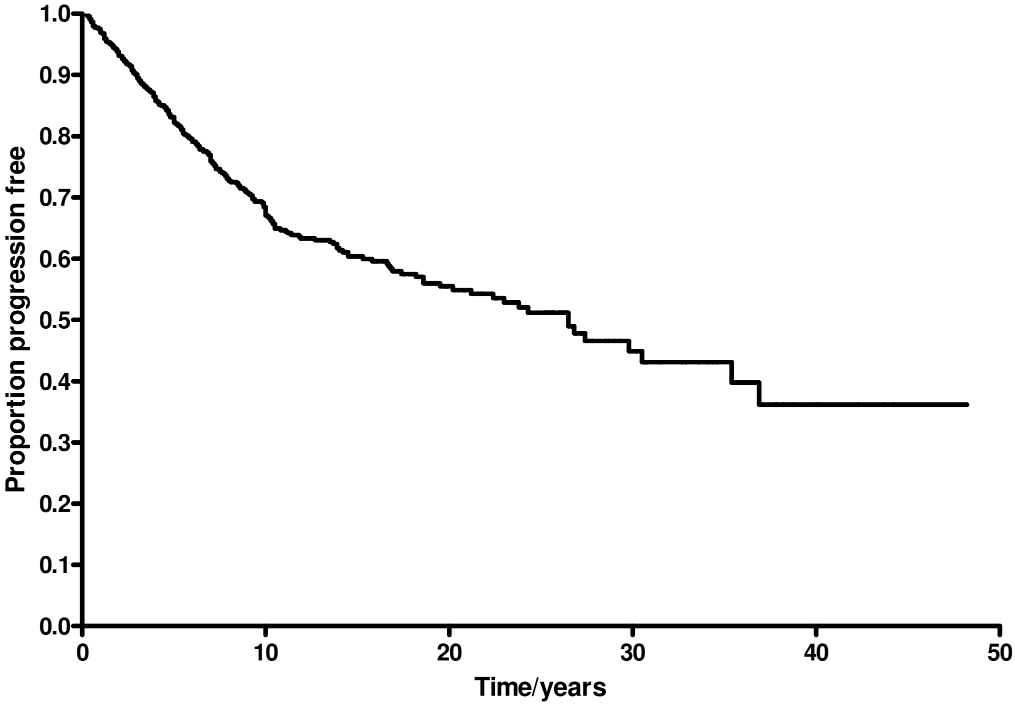


Figure 3-2 Kaplan-Meier of development of stricturing or penetrating disease, excluding those with disease progression at diagnosis

Time/years	% with stricturing or penetrating disease
1	3.1
3	10.5
5	17.8
10	32.9
20	44.5

Table 3-6 Percentages of patients with development of stricturing or penetrating disease, excluding those with disease progression at diagnosis

Rather than just plotting disease complications as a whole, the development of stricturing and penetrating disease was separated out (Figure 1-3). Patients who developed penetrating disease directly from inflammatory disease were assumed to have stricturing disease at the time of development of penetrating disease, as per the Montreal classification and as discussed in section 3.1. The median time to development of stricturing disease was 14.5 years whereas for penetrating disease this time was much higher at 52.2 years. At diagnosis, 18.0% and 8.9% of patients had already developed stricturing and penetrating disease respectively, and by 5 years these figures were 32.6% and 17.8% respectively (Table 3-7). Over time, the proportion of patients with disease progression increased: at 20 years 54.5% and 31.0% had developed stricturing and penetrating disease respectively.

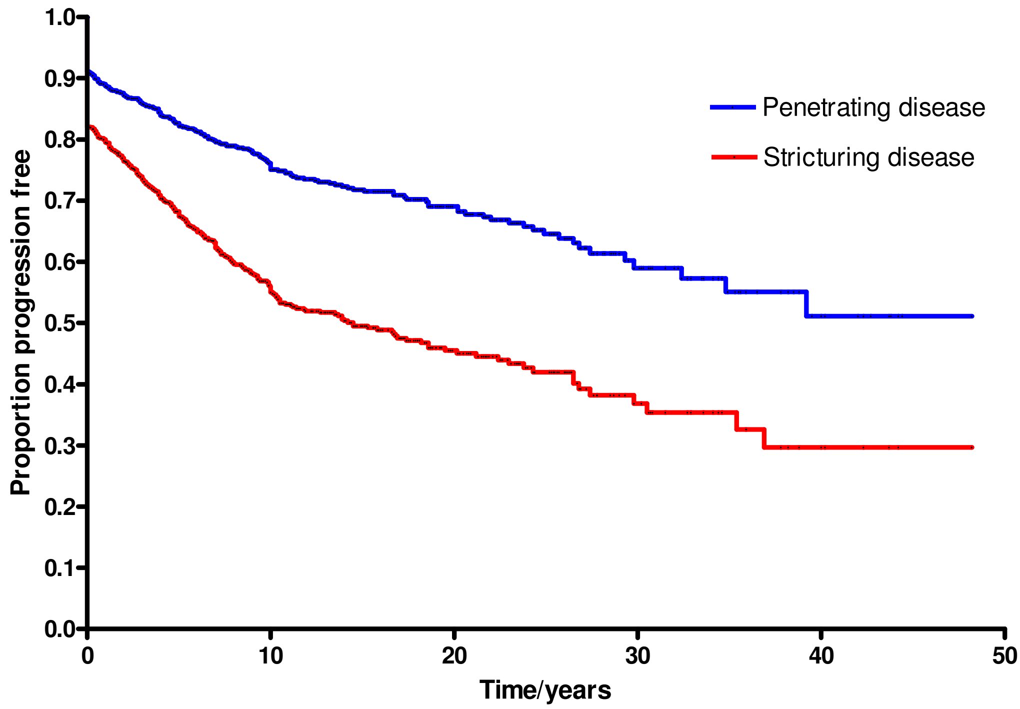


Figure 3-3 Kaplan-Meier curve of development of stricturing and penetrating disease

Time/years	% with progression	
	Stricturing	Penetrating
0	18.0	8.9
1	20.5	11.3
3	26.5	14.2
5	32.6	17.8
10	45.0	24.9
20	54.5	31.0

Table 3-7 Percentages of patients with development of stricturing and penetrating disease, including those with disease progression at diagnosis

3.5.2 Comparison of time to disease progression in Dundee and Edinburgh cohorts

Given the phenotypic differences in the Dundee and Edinburgh cohorts (Table 3-2) the time to disease progression was compared between the two cohorts. There was no statistically significant difference between the cohorts (log rank test p=0.378). The median time to disease progression for the Dundee and Edinburgh cohorts was 18.6 and 12.7 years respectively.

3.5.3 Time to disease progression according to disease location

To determine whether disease location affected time to progression, Kaplan-Meier survival analysis was used to compare disease progression in patients with different disease locations.

The initial analysis included all patients who had stricturing or penetrating disease at diagnosis (Figure 3-4). The median time to disease progression for L1, L2, L3 and L4 disease was 3.3 years, 36.9 years, 10.4 years and 5 years respectively (log rank test p <0.0001). As shown in Table 3-8, at 5 years 53.9% of L1 patients and 53.6% of L4 patients, compared with 11.5% of L2 and 33.3% of L3 patients had demonstrated disease progression. At 20 years the difference between L3 and L1/L4 disease was less marked, as 75% of L1, 84.9% of L4 and 65.7% of L3 patients had disease progression, although there was a clear difference between all of these and L2 disease in whom 20.4% of patients had progressed.

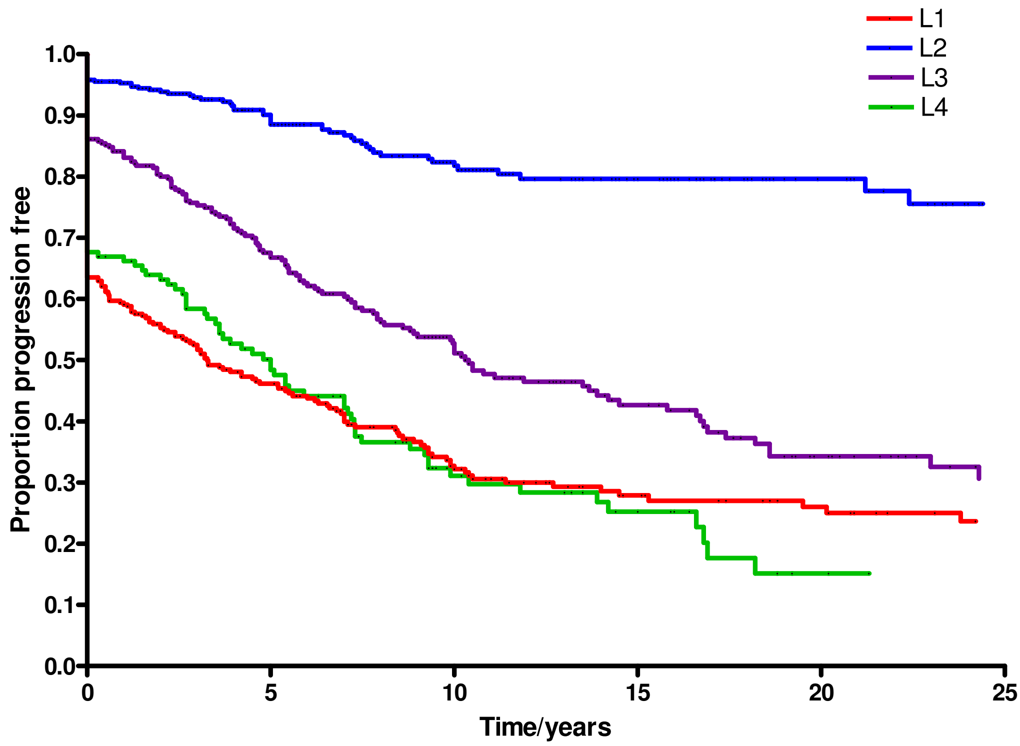


Figure 3-4 Kaplan-Meier of disease progression by disease location

Time/years	% with progression			
	L1	L2	L3	L4
0	36.5	4.2	13.9	32.4
1	40.9	4.8	16.9	33.8
3	48.3	7.3	24.7	41.6
5	53.9	11.5	33.3	51.6
10	67.8	18.3	48.9	68.9
20	75.0	20.4	65.7	84.9

Table 3-8 Percentages of patients with disease progression by disease location, including those with disease progression at diagnosis

As Table 3-8 shows, different proportions of L1 (36.5%) and L2 (4.2%) patients had disease progression at diagnosis. To determine the time to development of penetrating or stricturing disease for patients with inflammatory disease at diagnosis, all patients with stricturing or internally penetrating disease at diagnosis were removed and the analysis repeated (Figure 3-5). The median time to disease progression for L1, L2, L3 and L4 disease was 10.3 years, 36.9 years, 14.5 years and 9.3 years respectively (log rank test $p<0.0001$). As shown in Table 3-9, similar proportions of L1, L3 and L4 patients had disease progression at 5, 10 and 20 years, with L2 disease being markedly lower. When L2 disease was excluded there was no difference between the other groups (log rank test $p=0.121$).

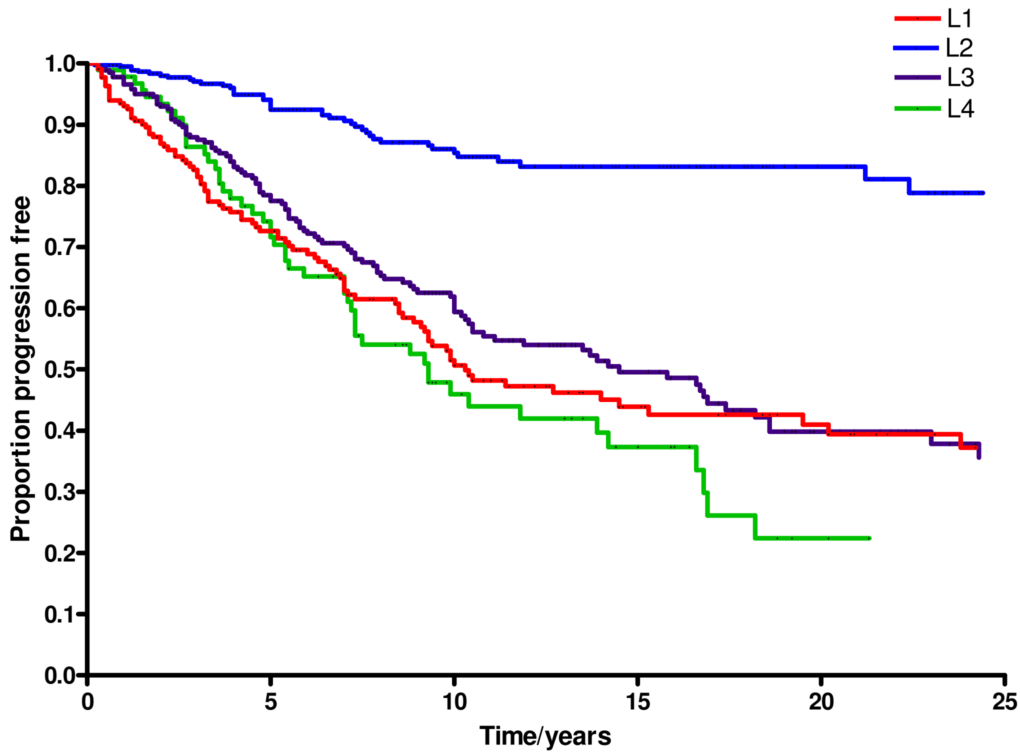


Figure 3-5 Kaplan-Meier curve of disease progression by disease location, excluding those with disease progression at diagnosis

Time/years	% with progression			
	L1	L2	L3	L4
1	7.0	0.5	3.5	2.2
3	18.6	3.0	12.5	3.7
5	27.4	7.6	22.5	28.4
10	49.3	14.6	40.6	54.0
20	59.0	16.9	60.2	77.6

Table 3-9 Percentages of patients with disease progression by disease location, excluding those with disease progression at diagnosis

3.5.4 Time to disease progression according to age at diagnosis

A Kaplan-Meier survival analysis was used to examine whether time to disease progression differed according to the Montreal age group at diagnosis. Although the median time to disease progression did increase across the age groups (A1, A2 and

A3 median times to disease progression 11.1, 14.2 and 16.7 years respectively), this did not reach statistical significance (log rank test p=0.8114).

3.5.5 Time to disease progression according to smoking at diagnosis

Kaplan-Meier survival analysis was used to examine whether time to disease progression differed according to smoking status at diagnosis (non-smoker, ex-smoker, current smoker). There was a statistically significant difference between the curves (log rank test p=0.011; median times to disease progression 18.2, 19.5 and 10.3 years for non-, ex- and current smokers respectively). This difference can be explained by the higher rate of disease progression at diagnosis (15.4%, 13.7% and 21.8% for non-, ex-, and current smokers respectively); when these patients were excluded from the analysis, there was no difference in median times to disease progression (29.8, 35.4 and 21.2 years for non-, ex- and current smokers respectively, log rank test p=0.1883).

3.5.6 Time to disease progression according to presence of perianal disease

Kaplan-Meier analysis was used to examine whether time to disease progression differed according to the presence of perianal disease at any point in disease course. There was no difference in median time to disease progression between patients with and without perianal disease (14.5 years for both groups, log rank test p=0.6544).

3.5.7 Time to disease progression - multivariate analysis

To determine the relative contribution of different factors to the risk of disease progression, a multivariate analysis was completed by Dr Nicholas Lewin-Koh at Genentech, SF using the R programme (version 2.10.1). A Cox proportional hazards model was used to examine the factors that significantly affect time to disease progression. The model included disease location, gender, age at diagnosis (as a continuous variable), smoking at diagnosis and the presence of perianal disease. Results showed that disease progression was not associated with gender, age at diagnosis, smoking at diagnosis or perianal disease. When compared to the risk of disease progression in patients with L2 disease, L1 disease conferred a hazards ratio (HR) of 4.7 (95% CI 3.6-6.1) and L3 disease conferred a HR of 2.8 (95% CI 2.1-3.8).

3.6 Surgical resection data

3.6.1 Time to first resection

The median time to first resection was 8.9 years (95% CI 7.5-10.0 years, Figure 3-6). A substantial proportion (12.7%) required a resection at diagnosis, as shown in Table 3-10. By 3 years, 32.8% of patients had required at least one resection, rising to 53.3% at 10 years and 68.9% at 20 years.

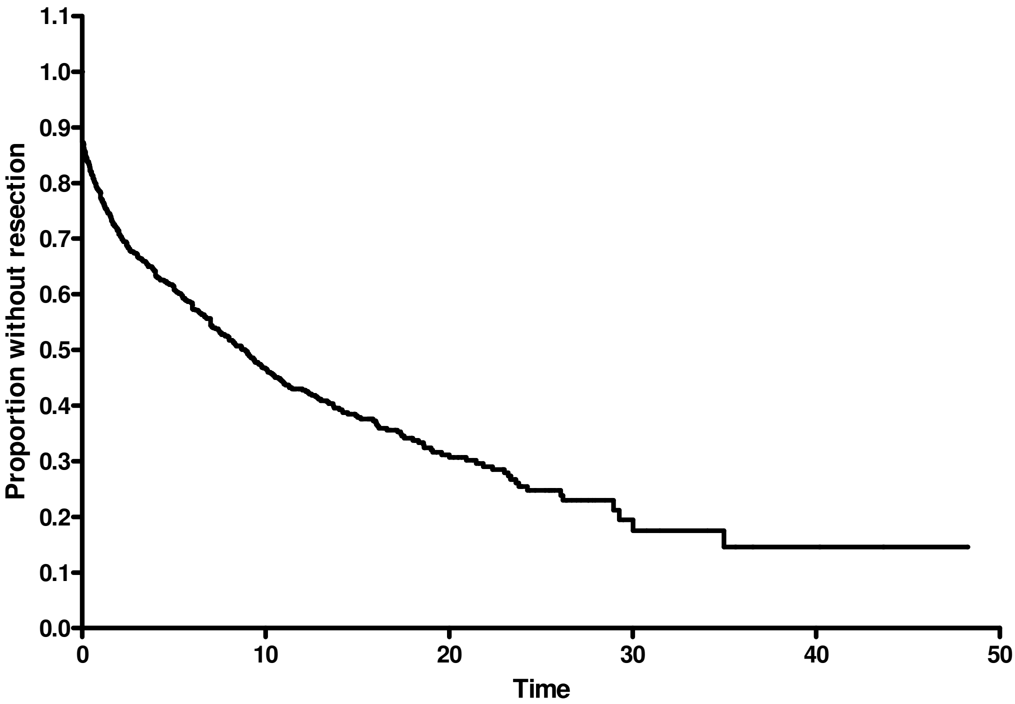


Figure 3-6 Kaplan-Meier curve of time to first surgical resection

Time/years	% having had 1st resection
0	12.7
1	22.7
3	32.8
5	38.6
10	53.3
20	68.9

Table 3-10 Time to first resection

3.6.2 Comparison of time to first resection in Dundee and Edinburgh cohorts

The time to first resection was compared between the two cohorts. There was no significant difference between the cohorts, but there was a trend for the time to first resection to be shorter in the Edinburgh cohort than the Dundee cohort (median times 8.0 and 10.8 years respectively, log rank test $p=0.0714$).

3.6.3 Time to first resection according to disease location

Kaplan-Meier analysis was completed for time to first resection according to disease location (Figure 3-7). The median time to 1st resection for L1, L2, L3 and L4 disease was 1.3 years, 30.0 years, 7.8 years and 7.0 years respectively (log rank test $p<0.0001$). As shown in Table 3-11, patients with L1 disease were particularly likely to have a resection at diagnosis, with 26.9% of L1 patients falling into this category, as opposed to 3.4%, 6.8% and 9.5% for L2, L3 and L4 disease respectively. The L1 disease location category had the lowest resection-free proportion at all points during follow up. Patients with L2 disease had a much lower risk of a resection than any of the other locations, and even at 20 years only 39.7% had needed at least one resection.

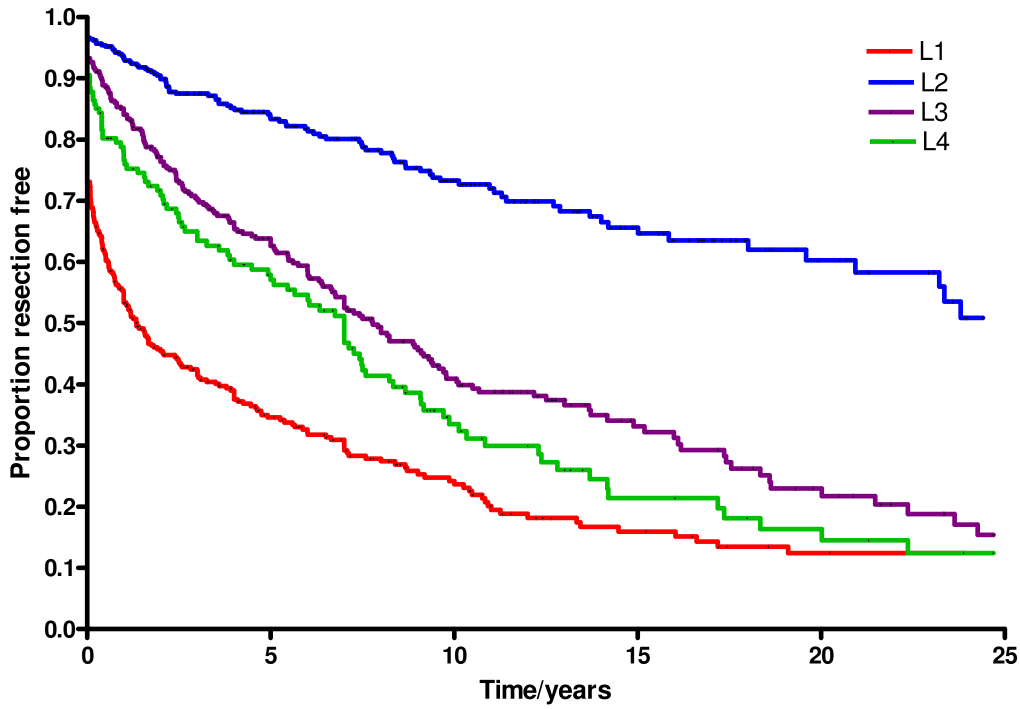


Figure 3-7 Time to first resection according to disease location, including those operated on at diagnosis

Time/years	% required first resection			
	L1	L2	L3	L4
0	26.9	3.4	6.8	9.5
1	46.6	6.9	15.9	23.3
3	57.9	12.5	29.6	36.6
5	65.4	16.3	37.4	49.9
10	76.4	26.8	59.1	66.5
20	87.6	39.7	78.3	85.5

Table 3-11 Percentages of patients needing resection by disease location, including patients operated on at diagnosis

The Kaplan-Meier analysis was repeated excluding patients who were operated on at diagnosis, and the statistically significant difference between the groups remained (log rank $p<0.0001$, Figure 1-8), even when L2 patients were not included in the analysis ($p=0.0002$). The median times to first resection for L1, L2, L3 and L4 disease were 4.5 years, 30.0 years, 8.9 years and 7.3 years respectively. As shown in

Table 3-12, the L1 disease location had the largest proportion of patients requiring resection within the first 5 years (52.7%), but by 10 and 20 years the differences between patients with L1, L3 and L4 disease were much less marked. L2 disease continued to have a lower risk of resection, with only 37.9% of patients followed up at 20 years having needed a resection.

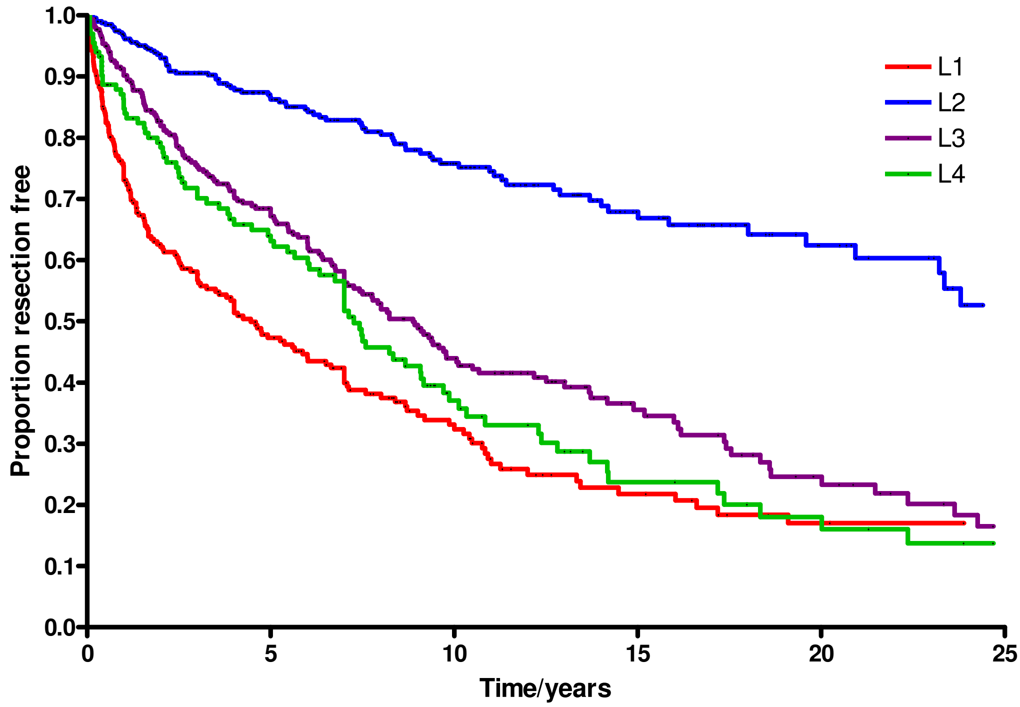


Figure 3-8 Time to first resection, excluding patients operated on at diagnosis

Time/years	% required first resection			
	L1	L2	L3	L4
1	27.0	3.6	9.8	15.3
3	42.3	9.5	24.5	29.9
5	52.7	13.4	32.8	36.9
10	67.6	24.2	56.1	63.0
20	83.0	37.9	76.7	84.0

Table 3-12 Percentages of patients needing resection by disease location, excluding patients operated on at diagnosis

3.6.4 Time to first resection according to age at diagnosis

Kaplan-Meier survival analysis was used to examine time to first resection according to the Montreal age group at diagnosis. There was no significant difference in the median time to first resection between the age groups (A1, A2 and A3: 9.0, 8.2 and 16.2 years respectively, log rank test p=0.228).

3.6.5 Time to first resection according to smoking at diagnosis

Kaplan-Meier survival analysis was used to examine time to first resection according to the smoking status at diagnosis (non-smoker, ex-smoker, current smoker). The median times to resection were significantly different (9.6, 19.1 and 6.5 years for non, ex and current smokers at diagnosis respectively, log rank test p<0.0001). This significance remained even when those operated on at diagnosis were excluded (p=0.0018, median times to resection 11.0, 20.9 and 9.7 years for non, ex and current smokers at diagnosis respectively, Figure 3-9). When ex-smokers were excluded, the differences between non- smoker and current smoker Kaplan-Meier curves did not quite attain significance (log rank test p=0.0537). The difference is not explained by disease location, as ex smokers had a statistically significant excess of L2 disease.

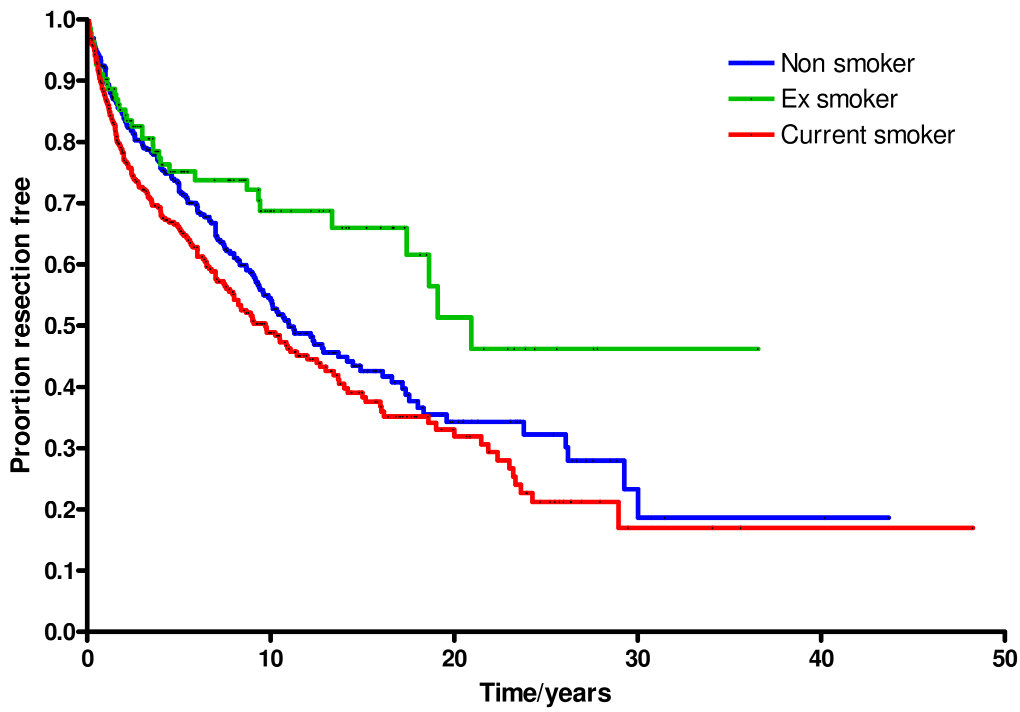


Figure 3-9 Time to first resection according to smoking status at diagnosis

3.6.6 Time to first resection according to presence of perianal disease

Kaplan-Meier survival analysis was used to examine whether time to first resection differed according to the presence of perianal disease at any point in the disease course. The median times to first resection for patients with and without perianal disease were 7.0 years and 9.4 years respectively but this was not significantly different (log rank test p=0.4185).

3.6.7 Time to first resection according to decade at diagnosis

As surgical practices have changed over the last few decades, a Kaplan-Meier analysis was completed comparing decade of diagnosis and risk of surgery. The numbers of patients diagnosed in each decade is shown in Table 3-13. The Kaplan-Meier graph is shown in Figure 3-10. The median times for those diagnosed in the 1950-60s, 1970s and 1980s were similar at 6.0, 6.9 and 7.1 years respectively, but for those diagnosed in the 1990s the median time to first resection was 10.1 years. For those patients diagnosed in the 2000s the median time could not be defined due to insufficient follow up. On a log rank test there was a statistically significant difference in the Kaplan-Meier curves between the groups (p=0.008) which became stronger on a log rank test for trend across the decades (p-value 0.0004), indicating an increase in the median time to first resection across the decades.

Decade at Diagnosis	Number of patients
1950-1960s	51
1970s	102
1980s	211
1990s	368
2000s	420
Excluded as surgical information not known	3
Total	1155

Table 3-13 Number of patients diagnosed in each decade

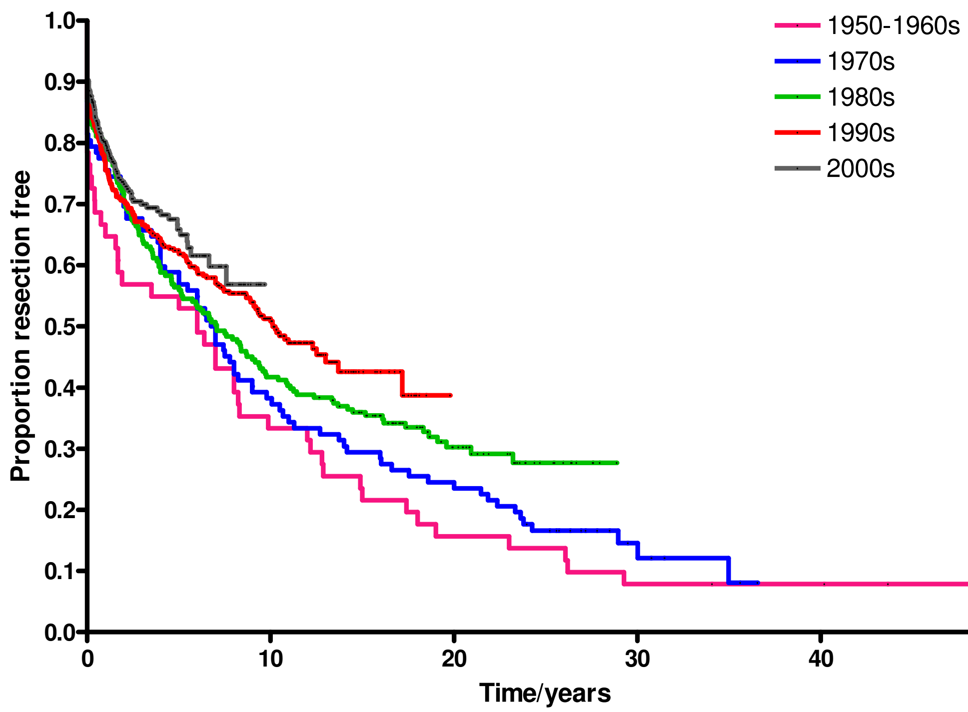


Figure 3-10 Kaplan-Meier curve of time to 1st resection by decade at diagnosis

3.6.8 Time to first resection - multivariate analysis

A Cox proportional hazards model was generated by Dr Nicholas Lewin-Koh to determine the factors that are independently associated with the risk of 1st resection. The model included gender, age at diagnosis (as a continuous variable), smoking at diagnosis and the presence of perianal disease. There was no association with age at diagnosis, gender, smoking at diagnosis or perianal disease. When compared with the risk of first resection for patients with L2 disease, L1 location conferred a HR of 5.2 (95% CI 4.1-6.5) and L3 disease conferred a HR of 2.6 (95% CI 2.1-3.3).

3.6.9 Multiple resections

The notes from the Scottish cohort were examined and information about dates of operations extracted. The maximum number of resections in any patient was seven. This multiple resection data was plotted on a Kaplan-Meier curve (Figure 3-11). For the total population the median time from diagnosis to the 2nd resection was 40.1 years, but for the 3rd-7th resections could not be defined as the majority of patients were not at risk, having not having had a 2nd resection.

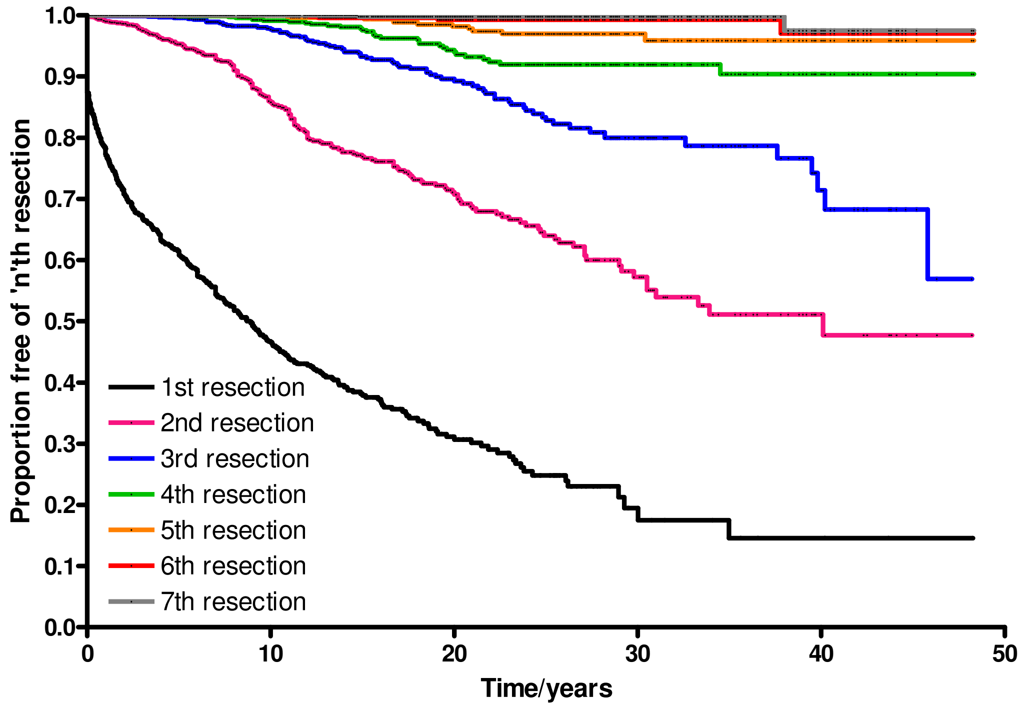


Figure 3-11 Kaplan-Meier curve of time to resection from diagnosis

The median time from diagnosis to the ‘n’th resection was calculated by considering only those at risk (Table 3-14). The median time from diagnosis to 2nd resection was long at 25.7 years and thereafter was above 30 years with 34.5 years being the median time from diagnosis to the 4th resection. Beyond the 4th resection the number of patients at risk was substantially reduced, making meaningful comparisons more difficult.

	Number at risk	Number having with date known	Had n^{th} resection but date not known	Number not having	Median time from diagnosis to n^{th} resection of those at risk	Median time from previous resection to next resection
1st resection	1155	589	1	565	8.9	
2nd resection	590	204	6	380	25.7	18.4
3rd resection	210	75	5	130	32.6	16.8
4th resection	80	31	4	45	34.5	9.2
5th resection	35	11	1	23	Undefined	Undefined
6th resection	12	4	1	7	37.8	16.8
7th resection	5	2	0	3	38.0	26.6

Table 3-14 Population at risk and median times to resection

The data were also analysed according to time from previous resection, arguably a more clinically relevant comparison. The median times from previous resection to the next resection are shown in Table 3-14. The Kaplan-Meier curve is shown in Figure 3-12. There was a statistically significant difference between the curves (log rank test, $p<0.0001$) and in addition the log rank test for trend between the 1^{st} and 7^{th} resection was also statistically significant (p-value <0.0001) indicating a trend for an increasing time between subsequent resections.

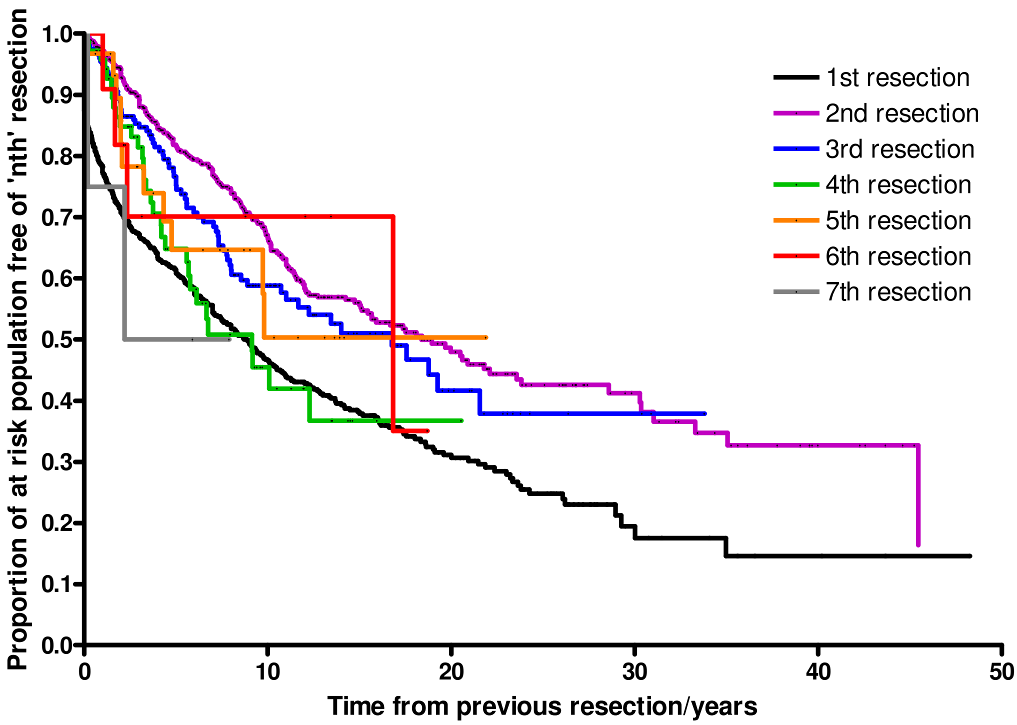


Figure 3-12 Kaplan-Meier curve of time from previous resection to next resection

3.6.9.1 Multiple resection data according to disease location

The multiple resection data were subdivided according to disease location. Only the first four resections were considered as subsequent resections had insufficient patients at risk in each disease location for the analysis to be meaningful. Times from each resection to subsequent resection were compared across each disease location in a Kaplan-Meier analysis; the results are shown in Table 3-15. L1, L3 and L4 locations all showed significant differences between times to subsequent resections, which was not seen in L2 disease. For L1 and L3 disease there was a significant decrease in the time intervals between subsequent resections, shown by the log rank test for trend (<0.0001 and 0.0097 respectively).

		1st resection	2nd resection	3rd resection	4th resection	Log rank p-value	Log rank test for trend
L1	Number at risk	360	264	86	25		
	Median time to resection/years	1.3	17.4	16.8	6.8	<0.0001	<0.0001
L2	Number at risk	416	101	22	6		
	Median time to resection/years	30	Undefined	Undefined	12.3	0.6021	0.9667
L3	Number at risk	325	186	75	32		
	Median time to resection/years	7.8	18.4	11.7	6.1	<0.0001	0.0097
L4	Number at risk	147	101	52	22		
	Median time to resection/years	7	10.7	7.2	6.8	0.0022	0.0580

Table 3-15 Disease location number at risk, median time to resection from previous resection and log rank test results

3.7 Detailed analysis of disease progression in the Dundee cohort

The Dundee cohort was examined in more detail to determine whether stricturing and internally penetrating disease are linked and should be considered sequential, or if they should be considered as completely separate markers of disease progression. The aim was to examine the patients who developed stricturing disease to determine whether they ever developed penetrating disease, and those who went from inflammatory to penetrating disease without an intermediate diagnosis of stricturing disease to determine whether any of them did have strictures at the time of diagnosis of penetrating disease, or if they subsequently developed strictures.

The patients from the Dundee cohort who had complete follow up documented throughout their disease course were considered (n=345 patients). There were 17 patients who had both stricturing and penetrating disease together as the first evidence of disease progression and were excluded from the following analyses. Patients who developed stricturing disease as the first evidence of disease progression were examined (n=84) using Kaplan-Meier analysis. The median time to development of penetrating disease was 32.4 years (Figure 3-13).

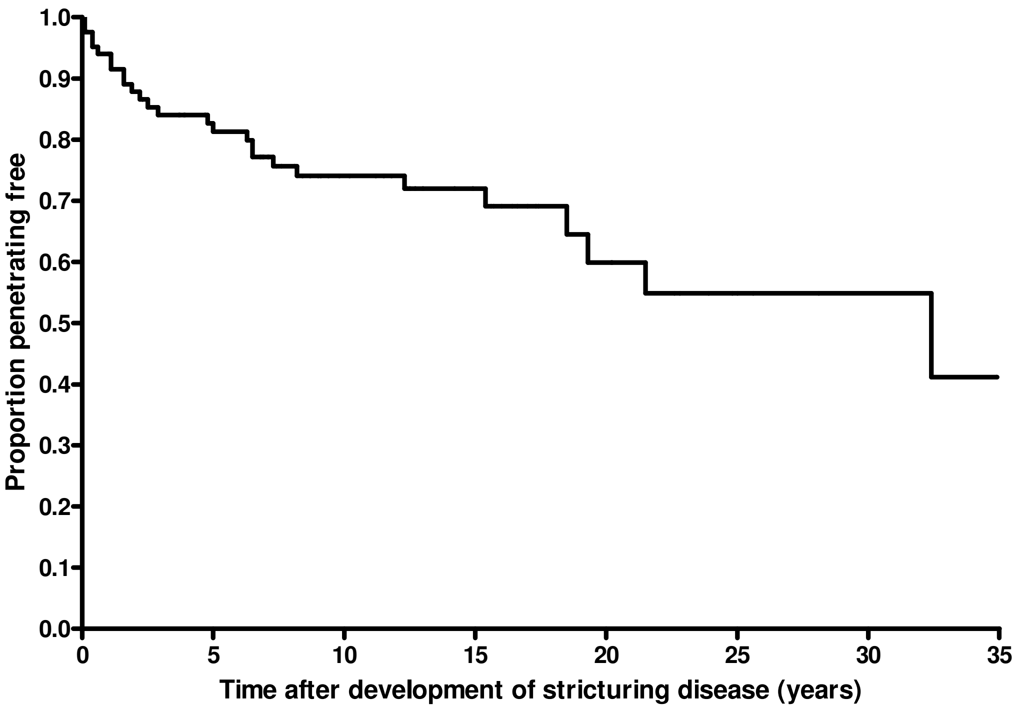


Figure 3-13 Kaplan-Meier curve of time to development of penetrating disease from time of diagnosis of stricturing disease

Similarly, patients who developed penetrating disease as the first evidence of disease progression were examined for development of stricturing disease (n=32, Figure 3-14). The median time to development of stricturing disease could not be defined, as 50% of patients had not developed stricturing disease during the follow-up period. A table showing the percentages of patients developing a different type of disease progression is shown in Table 3-16. Although the numbers of patients at risk is lower in patients with B3 disease as their first evidence of disease progression, very similar proportions of patients subsequently develop a different type of disease progression. At 10 years 25.9% of stricturing patients have penetrating disease, compared with 28.2% of penetrating patients have stricturing disease. On a log rank test there was no statistically significant difference between the two groups (p 0.843).

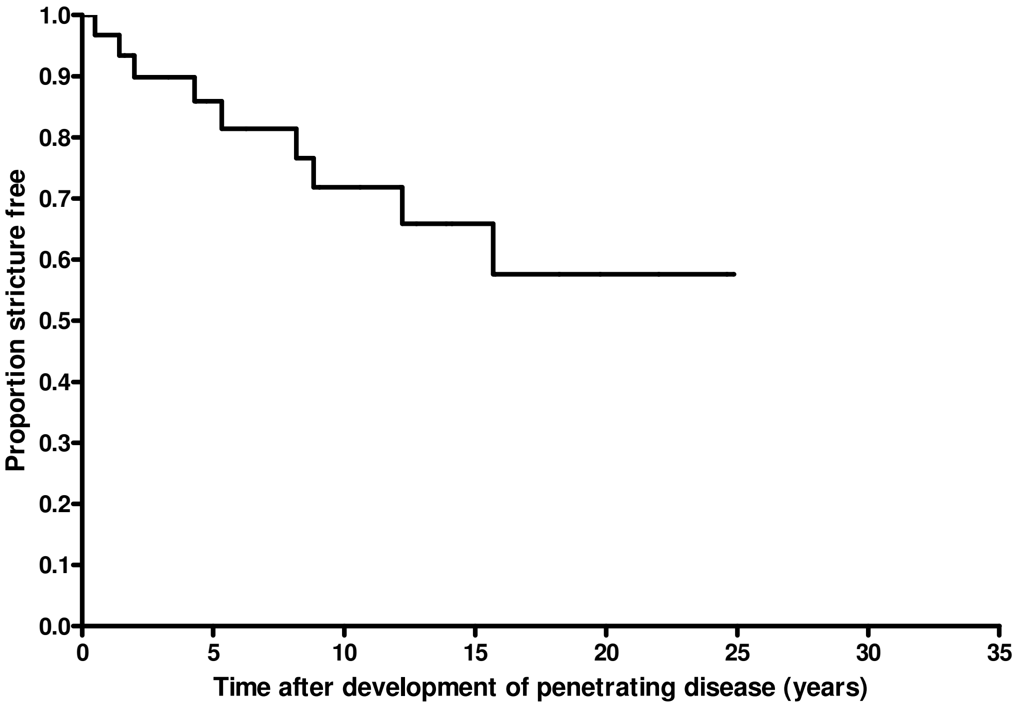


Figure 3-14 Kaplan-Meier curve of time to development of stricturing disease from time of diagnosis of penetrating disease

Time from initial disease progression/years	Stricturing-Penetrating (B2-B3) % with B2 disease developing B3 disease	Penetrating-Stricturing (B3-B2) % with B3 disease developing B2 disease
5	18.7	14.1
10	25.9	28.2
20	40.1	42.4

Table 3-16 Disease progression changes in the Dundee cohort

3.8 Discussion

3.8.1 Patient demographics

There were statistically significant differences in demographics and phenotype between the Dundee and Edinburgh cohorts. Whether there really are differences between cohorts which are geographically so close is unlikely. The higher proportion of young onset patients in the Edinburgh cohort is likely to reflect its position as a tertiary referral centre. The differences in disease location between the cohorts, especially in the proportions of L1 and L3 disease, as shown in Table 3-2, could possibly be because the two centres have different investigation algorithms; for example the Dundee cohort could do more routine colonoscopies of patients with ileal CD, thus finding more colonic disease or completing more small bowel investigations in those with colonic disease. Although there was a statistically significant difference in disease behaviour at 5 years between the two cohorts, as shown in Table 3-2, this was not reflected in differences in disease progression when the two cohorts were compared in a Kaplan-Meier analysis in section 3.5.2, and is therefore unlikely to be a real difference.

Smoking status markedly affected disease location, with patients who were smoking at diagnosis much more likely to have ileal disease and less likely to have pure colonic disease than non- or ex-smokers. This is in keeping other published cohorts.[221;222]

3.8.2 Comparison of disease progression in Scotland with other cohorts

Two of the largest patient cohorts studied for disease progression are from France[223] (Cosnes *et al.*) and New Zealand[224] (Tarrant *et al.*). A table comparing the basic characteristics of the cohorts is shown in Table 3-17.

Country	Pt no.	Follow up	Age at diagnosis (%)			Location (according to Vienna classification[17]) % of patients in each group				
			A1	A2	A3	Time of assessment	L1	L2	L3	L4
Scotland	1155	Median 10.3 yrs (IQR 4.1-19.7)	10	61	29	Maximum extent	27	34	24	13
France[223]	2002	Mean 8.6 +/-7.9 years	Not given			Maximum extent	Unclear			
New Zealand[224]	715	Median 6.5 years	12	58	30	At diagnosis	32	49	19	1

Table 3-17 Comparison of basic demographics

The French dataset, although much bigger than the Scottish cohort (2002 patients compared with 1155), had similar numbers of patients still under follow up at 20 years: approximately 300 (15%) compared with 283 patients (24.5%) in the Scottish cohort. The French data were analysed according to the Vienna classification; thus the definitions of location were slightly different and perianal fistulae counted as penetrating disease. Due to the way the data were collected in the Scottish cohort, it was not possible to analyse the data according to the Vienna classification of disease behaviour, making a direct comparison difficult. In the French cohort, perianal disease accounted for 540 (57%) of the 945 penetrating disease progressions. Of these, in 433 (80%) no internally penetrating behaviour was ever observed; it was only in the remaining 107 patients (20%) that an internally penetrating complication followed. This indicates that out of the total population, although there were 945 penetrating disease complications, only 512 (54%) ever developed internally penetrating disease. Therefore it is not surprising that the figures for development of disease progression were much higher compared with the Scottish cohort (Table 3-18), with the risk at 20 years of having stricturing or penetrating disease being 88%

in the French cohort, compared with 54.5% in the Scottish cohort. Risk factors for penetrating disease in the French cohort were age at diagnosis <40 (HR 1.4, 95% CI 1-1.5), being non Causcasian (HR 1.3, 95% CI 1.1-1.6), anoperianal lesions (HR 2.6, 95% CI 2.3-3.0) and lack of oesophagogastroduodenal involvement (HR 1.4, 95% CI 1.1-1.9). Risk factors for stricturing disease included ileal or jejunal involvement and lack of colonic involvement (HR 2.5, 3.2 and 2.0 respectively) as well as lack of anoperineal involvement (HR 1.4). For the disease location variable it is not clear what the reference was for the hazard ratios. The multivariate analysis data presented in this chapter was on the basis of time to first evidence of disease progression, rather than considering stricturing and penetrating disease separately; disease location also came out strongly, but age at diagnosis was not significant, nor was perianal disease. Perianal ('anoperineal') being a risk factor for penetrating disease yet protective against stricturing in the French dataset can be explained by their data collection. In their population, 46% of the patients who developed penetrating disease only ever had perianal disease as their manifestation of penetrating disease; this would explain why perianal disease was a risk factor for penetrating disease. Anoperineal disease 'protecting' against stricturing disease would have been because patients directly developing penetrating disease were excluded from further analyses of stricturing disease, having been 'censored' for stricturing disease at the time of penetrating disease development.

In the New Zealand cohort, where disease behaviour was phenotyped according to the Montreal classification, the risk of disease progression was much higher than in the Scottish cohort (45% vs. 17.8% at 5 years in those patients with inflammatory disease at diagnosis), as shown in Table 3-18. The median time to disease progression was also shorter (21.2 years compared with 26.5 years in the Scottish cohort, only considering those with B1 disease at diagnosis). The presence of perianal disease was predictive of disease progression (HR 1.62, 95% CI 1.28-2.05), as was L3 and L1 disease (HR not stated in the paper).

Cohort		% Risk of progression at different time points				Risk factors for disease progression
		0	5yrs	10yrs	20yrs	
Scotland	Including B2/B3 at Dx	18.0	32.6	45.0	54.5	Ileal/Ileocolonic disease
	Excluding B2/B3 at Dx	0	17.8	32.9	44.5	
France	Including B2/B3 at Dx	NG	52.0	NG	88.0	Age, Absence of colonic disease, Perianal disease
New Zealand	Excluding B2/B3 at Dx	0	45.0	56.0	NG	Ileal/Ileocolonic disease, Perianal disease

Table 3-18 Risk of disease progression in the different cohorts, NG= Not given

The comparisons suggest that the Scottish phenotype is less severe than that in New Zealand. This is especially surprising because the New Zealand cohort was population-based, as opposed to the clinic-based Scottish cohort: it is normally expected that a clinic-based study has a greater chance of being skewed in the direction of those with more severe disease. It is possible that CD is managed more aggressively in Scotland, which may prevent disease progression. It is also possible that Scottish patients were diagnosed earlier in the course of their symptoms. However, the differences in disease progression are so large that it seems unlikely that either of these two possibilities could adequately explain them. Alternatively there could be due to differences in interpretation of the Montreal classification, e.g. a mild inflammatory stricture without prestenotic dilatation and without obstructive symptoms could have been interpreted as B2 in the New Zealand cohort, whereas in the Scottish cohort this would be interpreted as B1. Without cross-validation of the New Zealand phenotyping this question is impossible to answer.

The consistent risk factor for disease progression across all three cohorts was disease location, especially L1 disease. In the French cohort L1 disease conferred a HR of 2.5 (95% CI 1.9-3.3) for B2 disease, compared to 5.2 (95% CI 4.1-6.5) for disease progression (either B2 or B3 disease) in the Scottish cohort. Unfortunately the New Zealand paper did not state HR for L1 and L3 disease. It is not understood why ileal disease is such a risk factor for disease progression. It has previously been argued that the narrow ileal lumen compared with the colon might predispose to the development of ‘functionally significant’ strictures.[225] The reduced mobility of the colon compared with the small bowel could result in fewer opportunities for serosal surfaces to stick together. Alternatively, the larger surface area of the small bowel compared with the colon could increase the chances of fibrosis and fistulae formation, despite similar cell turnover times. Further work is required in this interesting area.

Perianal disease, both on univariate and multivariate analyses, was not associated with disease progression in the Scottish cohort, unlike the New Zealand cohort, despite both cohorts being phenotyped according to the Montreal classification. In the New Zealand cohort 27% of patients had perianal disease; this figure in the Scottish cohort was 22.5%. It is possible that the shorter follow up period in the New Zealand cohort could skew the data in favour of patients with perianal disease.

3.8.3 Risk of resection

The median time to first resection in the Scottish cohort was 8.9 years, with an association with disease location and risk of resection (HR 5.2 for L1 disease compared with L2 disease and 2.6 for L3 disease compared with L2 disease). It is unfortunate that the multivariate analysis did not take account of decade at diagnosis, as this was the only other variable associated with risk of surgery on univariate analyses, and earlier decade at diagnosis has been associated with a shorter time to operation in another cohort.[226]

Country	Patient no	Median Follow up length (yrs)	Median age at Dx	Age at Diagnosis (%)			Location (%)				
				A1	A2	A3	Time point of assessment	L1	L2	L3	L4
Scotland	1155	10.3	30.8	10	61	29	Vienna Max extent	27	34	24	13
USA	345	Not given	25.8	34.7 (<20 yr)	35 (20-40 yrs)	41	Vienna at diagnosis	36	33	30	5
Norway	237	Not given	Not given	70		30	Vienna at diagnosis	27	49	23	1
Wales	341	7.7	30	9	57	34	Montreal at diagnosis	38	45	14	4

Table 3-19 Basic demographics

Cohort	Definition of surgery
Scotland	Intestinal resection, excluded simple appendicectomies and defunctioning procedures
USA	Any intra-abdominal surgical procedure performed for treatment of CD or its complications
Norway	Any intra-abdominal surgical procedure for active CD
Wales	Resection of bowel, stricturoplasty or defunctioning stoma formation

Table 3-20 Definitions of surgery in different studies

Relatively few cohorts have examined the variables associated with risk of resection. Three cohorts compared here to the Scottish data are from USA[227], Norway[228] and Wales.[226] The basic demographics of the cohorts are compared in Table 3-19.

In the US cohort of 345 patients[227] diagnosed between 1991 and 1997 in New England, 24% of patients required 'major surgery' within 3 years of diagnosis, as shown in Table 3-21. The US definition of major surgery, as shown in Table 3-20, was less stringent than that applied in the Scottish study. Despite the less stringent definition, the risk of surgery in the US cohort was less than that in the Scottish cohort (24% vs 33% at 3 years). On univariate analysis, risk factors for major surgery within 3 years included being a current smoker (HR 3.1, 95% CI 1.47-6.51, compared with non-smokers) and L1 location (HR 2.22, 95% CI 1.30-3.81, compared with all other locations). L2 location was found to be protective against surgery (HR 0.27, 95% CI 0.13-0.56, compared with all other locations).

A smaller Norwegian cohort of 237 Crohn's patients[228] examined risk of surgery, with the definition similar to the US study detailed above, as shown in Table 3-20. By 5 years, 27% of patients had required surgery, compared with 39% in the Scottish cohort, as shown in Table 3-21.

	% Risk of surgery					
	0	3	5	10	20	Risk factors
Scotland	12.7	32.8	38.6	53.3	68.9	Ileal/ileocolonic location
USA	4.1	24.2	NG	NG	NG	Smoking, Ileal location, non colonic disease
Norway	0	NG	27	37.9	NG	Ileal location, stricturing or penetrating disease
Wales	7	NG	39	NG	NG	Year at diagnosis, colonic location (protective), early oral steroid therapy, early thiopurines use (proective)

Table 3-21 Risk of surgery in the different cohorts, NG= Not given

Both cohorts showed a lower rate of surgery compared with the Scottish cohort. Even if only those people diagnosed more recently in the Scottish cohort were considered, the risk of surgery was still substantially higher than in the other studies (37.9% and 34.2% of patients diagnosed in the 1990s and 2000s respectively in the Scottish cohort at 5 years, compared with 27% in the Norwegian study overall at 5 years, all of whom were diagnosed 1990-1994).

This is in contrast to the Welsh study[226], which, unlike the US and Norwegian studies, showed similar rates of surgery to the Scottish cohort. The Welsh study also noted a greater risk of surgery with those diagnosed in earlier decades (HR 1.7 95% CI 1.1-2.5 for those diagnosed 1986-1991 compared with those diagnosed 1998-2003), confirmed the finding of the other studies that colonic disease was protective (HR 0.39 (95% CI 0.26-0.56 compared with ileal disease), but also found that early oral steroids and early thiopurines use were also associated (HR 1.7 and 0.47 respectively).

The reasons behind the differences in risk of surgery between the US/Norway and the UK are not clear. One possibility is that the Scottish/Welsh CD phenotype is more severe although the risk of disease progression (a surrogate marker for severe disease) is lower in the Scottish cohort than in other cohorts. As will be discussed in Chapter 4, there is some evidence to suggest that when other criteria are used to assess severity of disease the Scottish phenotype does appear to behave in a more severe manner than some other cohorts. Patient recruitment (clinic-based system vs. population-based) could also account for the differences, skewing the data towards people who required more hospital intervention. As L1 disease is a risk factor for need for surgery, the higher rate of surgical intervention in the Scottish cohort could be explained if there was a higher rate of L1 disease, but this was not the case. Another possibility is a difference in surgical practices in Scotland compared with other countries, although this would be difficult to investigate or prove.

In conclusion, the higher rate of surgery in the Scottish and Welsh cohorts may be a reflection of a more severe disease phenotype, the clinic based nature of recruitment, and possibly a difference in surgical practices in the UK compared with other centres.

3.8.4 Multiple resection data

There was a trend for time from one resection to the next to lengthen gradually, with the median time to 1st resection being 8.9 years, compared with median times of 18.4 and 16.8 years between 1st and 2nd resections and 2nd and 3rd resections respectively. When this was broken down into disease locations, comparisons became more difficult because of smaller patient numbers. Patients with L4 disease appeared to have the shortest intervals between resections with little evidence for an increase in time to the next resection. Similarly with L1 and L3 disease there was also a trend for lengthening times between resections. Conversely, a trend was not proven for patients with L2 disease, who appear to have very long intervals between successive resections. It is to be expected that patients with L2 disease would have long intervals between operations: disease location tends to remain static over time[224], therefore patients requiring colectomy for L2 disease are much less likely to need a further resection than patients with other disease locations. Conversely, operations for L1 and L4 disease are likely to remove the minimum amount possible, meaning a greater chance of needing a further operation.

It would be interesting to analyse the data further to examine factors that may be important in the early need for re-operation. This was studied in a cohort of 432 CD patients in Milwaukee recruited 1998-2004[229], of whom 65 patients required more than 1 surgical intervention. Of these 65 patients, 32 (49%) required 'rapid reoperation' (i.e. repeat intestinal surgery, including stricturoplasty and resection, within 2 years of the previous operation. However, no significant differences were found between the rapid and non-rapid reoperation groups in terms of disease location, disease behaviour or age at onset. Another study in Pennsylvania examined 88 patients who had had at least 2 resections for CD, and whose last resection had been between the years of 1988 and 1993.[230] Risk factors for early post operative recurrence of CD after the 1st resection were 'perforating' disease (HR 2.33, 95% CI 1.47-3.57, compared with 'non perforating' disease) and a period of >8.9 years from the diagnosis of CD to the 1st resection (HR 1.96, 95% CI 1.06-3.61) compared with a preoperative period of <8.9 years). Disease location, age group at diagnosis and smoking history were not found to be significant. An earlier study in New York examining 164 patients who had an intestinal resection 1976-1989, had found that early symptomatic recurrence occurred in 40% of patients and was not associated with age at onset, or location of disease, but did appear to be associated with histological evidence of CD in the resection margins and the number of anastomoses (HR not given).

These studies suggest that disease location and age at diagnosis are not important factors in early reoperation, and that disease behaviour is important. Other factors must therefore be important, which merits further investigation.

3.8.5 Disease behaviour

Although the numbers of patients developing B3 disease as their first manifestation of disease progression ('primary B3 disease') is less than the numbers developing B2 disease as their first manifestation of disease progression ('primary B2 disease'), it is interesting to note that patients with primary B3 disease have an equal risk of subsequently developing stricturing disease as the risk of patients with primary B2 disease have of developing B3 disease. It would be helpful to extend this examination to the full Scottish cohort.

It must be noted, however, that this was a retrospective study and that detailed pathological examination of resection specimens, over and above what is routine clinical practice, was not completed. Decisions about B2 and B3 disease were made on the basis of what was documented in the clinical notes, including radiological, pathology and endoscopy reports, as ascertained by a clinician. Therefore it is hard to prove one way or another in this cohort whether B3 disease really does always coexist with B2 disease or not. A further prospective study examining pathology resection specimens to further elucidate the pattern of disease behaviour in CD would be necessary. Certainly, considering B2 and B3 as entirely separate entities would seem to be appropriate in the interim.

3.8.6 Conclusions

This study of disease behaviour, using carefully and consistently obtained phenotypic data, is one of the largest cohorts examined to date for disease progression and is the largest study examining risk of surgery in CD. It provides further confirmation of the importance of disease location in determining disease behaviour and the need for surgery. It also confirms that surgical practices have changed over the years, with patients being at a lower risk of surgery now compared with previous decades. The study provides novel perspectives on the need for multiple resections, showing that disease location affects the need for subsequent operations. It provides data on disease progression, suggesting that B2 and B3 behaviour should be considered as separate entities, something that has important implications for the Montreal classification and disease phenotype in general.

The limitations of the current study include the fact that it is a clinic-based rather than a population-based cohort study. In addition, it was retrospective rather than prospective. Prospective cohort studies are an important 'gold standard' in research methodology but have the limitation of taking many years to generate appropriate data. Further studies examining the effect of drug therapy on disease progression and need for surgery are necessary, especially to examine the hypothesis that early biological and immunosuppressive therapies change long term disease course rather than just preventing relapse.

However, despite these limitations, the results presented in this chapter demonstrate that important research can still be generated with retrospective studies which have important implications for the future management of patients with CD. As risk of disease progression and the need for surgery are surrogate markers for disease severity, alternative methods of defining disease severity are required. This concept is explored in more detail in the next chapter.

Chapter 4 Crohn's disease severity prediction

Summary

Aims: To construct a multi-faceted composite score that could be used to define a proportion of the population with severe disease, and to find clinical and genetic factors that correlated with more severe disease.

Methods: A novel severity score assessed over the first 5 years after diagnosis was constructed based on physician consensus, with a possible score that ranged from 1 to 16. This was applied to 366 CD patients recruited in Dundee. Univariate and multivariate analyses (using the chi squared test and logistic regression with backward elimination respectively) were completed to examine for factors present at diagnosis (clinical and genetic) that correlated with more severe disease, defined as a score >6. The 30 top susceptibility loci uncovered by a recent CD GWAS were genotyped in the Dundee cohort of 366 CD, 261 UC patients and 539 controls using the Sequenom® platform at the University of California, San Francisco. A case-control analysis was completed for these SNPs in both CD and UC.

Results: 249 CD patients had full severity data available for the first 5 years after diagnosis. The mean score was 6.3. Association with long term disease progression was demonstrated for a subset of the score. A statistically significant correlation between the proportion of risk alleles present and severity score was not found (R^2=0.012, p=0.078). On univariate analyses, age <17 at diagnosis (OR 0.23 compared with patients 17-40 at diagnosis, p-value 0.0004), ileal disease at diagnosis (OR 2.2, p-value 0.0025) and upper GI disease at diagnosis (OR 5.3, p-value 0.0008) correlated with more severe disease. Using logistic regression these three factors along with 2 SNPs (rs13361189 and rs9286879) were found to have an independent correlation with more severe disease, and were built into a model to define the probability of more severe disease. On a case-control analysis 8 SNPs (representing loci at or near *IL23R*, *ATG16L1*, *PTGER4*, *IRGM*, *ZNF365*, *NOD2* 1007fs, *ORMDL3* and the SNP rs1736135) showed association with CD susceptibility with similar OR to those in the original GWAS meta-analysis. Three meta-analysis SNPs (representing loci at or near *IL12B*, *STAT3* and *IL23R*) were associated with UC susceptibility. There was a statistically significant difference in the mean proportion of risk alleles present in CD compared with controls.

Conclusion: This study has demonstrated that it is possible to construct a composite severity score that is correlated with long term outcomes and which is more discriminant than existing definitions of CD severity. In addition, it has been used to build a model for predicting severe disease that could be useful in clinical practice. Further validation in a separate cohort is required.

4.1 Introduction

Recent treatment algorithms for CD are increasingly advocating a ‘top down’ approach to the management of the disease.[231] That is, giving immunosuppressant therapy with thiopurines and biologicals at an early stage with the hope of preventing the development of strictures and fistulae. However, long term immunosuppression is not without its risks, including infections, blood dyscrasias, and, with biologicals, possibly the development of autoimmune phenomenon, e.g. demyelination and lupus, as well as solid cancers.[232;233] Therefore it is important to select patients carefully on the basis of long-term risk of severe disease, to prevent giving people strong immunosuppression unnecessarily. Unfortunately there is no test or objective way to identify those who would benefit most from early therapy in CD. Ideally, the best criteria should be apparent at diagnosis so that an informed decision can be made about the patient’s long term risk of having severe disease at an early stage.

Currently there is no universally recognised definition of what constitutes severe or disabling disease. The ‘behaviour’ variable of the Montreal classification[20] is the closest to approximating the development of disabling disease, as people with penetrating disease are at higher risk for operations than those with inflammatory disease.[228] In addition, there is a clear progression over time from inflammatory disease to the development of stricturing and/or penetrating disease as demonstrated by other authors[223] and in the previous chapter.

The only current classification of severe disease is by Beaugerie *et al.*[234]; their criteria for severe disease are given in Table 4-1.

Within the 5 year period following diagnosis, at least one of:	
1.	More than 2 steroid courses required and/or dependence on steroids
2.	Further hospitalization after diagnosis for flare-up or complication of the disease
3.	Presence of a cumulative time of more than 12 months within the 5 year study period of disabling chronic symptoms, this being defined as including: -Diarrhoea with nocturnal and or urgent stools -Intense abdominal pain because of intestinal obstruction -Fever or fatigue attributable to the disease -Joint pain -Painful uveitis or pyoderma gangrenosum
4.	Need for immunosuppressive therapy
5.	Intestinal resection
6.	Surgical operation for perianal disease

Table 4-1 Beaugerie criteria for disabling Crohn's disease

In their cohort 85.2% of patients fitted this definition of disabling disease. In a cohort of Scandinavian patients[235], 60% fitted the criteria for 'disabling' disease. Thus it could be argued that this description is not very discriminating. The other issue with the Beaugerie criteria is the description of disabling disease itself. The need for surgical operation for perianal disease is the criterion that is the easiest to fulfill, as almost all patients who have perianal disease will need a minor operation, e.g. incision and drainage, or seton placement. It is therefore not surprising that on multivariate analyses perianal disease comes out as a strongly positive predictor of 'disabling disease'.

The hypotheses of this chapter were:

1) That a multi-faceted composite score could be used to define a proportion of the population most at risk of severe disease.

2) That phenotypic and genetic variants could be correlated with disease severity, based on the severity score, and that genetic variants, as an unchanging variable, could be used from the onset of disease to help predict disease severity, before complications developed.

4.2 Methods

4.2.1 Severity score

The severity score itself is detailed in the Chapter 2, and again in Table 4-2. It was developed in consultation with various members of the gastroenterology unit at the Western General Hospital, Edinburgh (Prof Jack Satsangi, Dr Gwo-Tzer Ho and Dr Ian Arnott) as well as Dr Craig Mowat at Ninewells Hospital, Dundee.

Using the Beaugerie definition of ‘disabling disease’ as a starting point, an initial discussion was had with each Edinburgh-based clinician as to what they thought defined CD severity. When it became apparent that there were several different aspects to defining severity, the consensus of opinion was that a composite score reflecting different disease parameters would be the best. Discussion between clinicians was had as to what the disease parameters should be. Disease phenotype and disease management encompassed disease progression and surgical management respectively; nutritional status accounted for poor nutrition as a result of more severe CD and hospitalisations reflected the social-economic impact, as well as recognising the fact that more difficult-to-manage CD is more likely to result in a hospital admission. Within each parameter, consensus was reached as to what severe, moderate and mild disease should be. Once overall consensus was reached between Edinburgh clinicians, it was sent to a clinician in another hospital (Dr Craig Mowat, Ninewells Hospital, Dundee) for independent verification that this score accurately reflected an expert’s opinion as to what constituted severe disease. Dr Mowat suggested some minor refinements, and these were accepted by all Edinburgh clinician. Following this, no further modifications were forthcoming from any of the Scottish IBD experts. Of note, BMI was used as the indicator of nutritional status and was not corrected for BMI percentile in those aged less than 17 at diagnosis. Although recognising that this was not optimal for those yet to finish puberty with epiphyseal fusion, patients less than 14 at diagnosis (an age at which 50% of children will have completed puberty) accounted for less than 5.7% of the cohort.

A time point of the first 5 years after diagnosis was chosen to assess this score which was felt to give the best balance of maximising patient numbers yet assessing severity on a sufficiently long timescale. The minimum total possible score was 1 with a maximum of 16.

The information was collected from all previously recruited Dundee CD patients by retrospective case note review, with the help of Dr Tim Heron in Ninewells Hospital, Dundee.

	Severe: Score 4 for each	Moderate: Score 2 for each	Mild: Score 1 for each	Score 0
Disease extent/behaviour	Panenteric disease OR Complex perianal disease requiring 3 or more operations OR Fistulating disease	Stricturing but not fistulation Perianal disease requiring 1 or 2 operations	Single site involvement No evidence of stricturing/fistulation Perianal disease but not needing operation	
Medical/surgical management	Steroid dependency OR Need for 2 or more immunomodulatory drugs OR 2 Surgical resections OR Use of biological therapy	More than 4 steroid courses, but none >4 months OR 1 immunomodulator OR 1 surgical resection	1-3 courses of steroids, each lasting <4 months No immunomodulators	No steroids No immunomodulators
Nutritional status	BMI<15 at any point in the 5 years	BMI=15-18.5 at any point in the 5 years		BMI >18.5 at all times in the 5 years
Socio-economic impact	5 or more hospitalizations for management of active disease	2-4 hospitalizations for active disease	1 hospitalization for active disease	No hospitalizations for active disease

Table 4-2 Severity score, calculated for the first 5 years after diagnosis

4.2.2 SNP selection

Thought was given to the best SNPs to select to genotype in the Dundee cohort. Most of the CD-related SNPs are associated with disease susceptibility rather than any other aspect of the disease, so it was thought that these would not be associated with disease severity in a given population. The initial plan had been to genotype the top 50 SNPs from the WTCCC GWAS study[73] that appeared to differentiate between B1 and B2/B3 disease at 5 years, and examine them with respect to the new severity score. This was discussed in detail with Dr Carl Anderson of the WTCCC, who stated that the most significant SNPs differentiating B1 and B2/B3 disease at 5 years in the CD dataset had not been replicated in a separate cohort (unpublished data). Despite this lack of replication it was still felt to be the best strategy, as SNP selection was on the basis of a surrogate marker of disease severity rather than susceptibility. The co-operation and permission of the UK IBD Genetics Consortium was obtained. Unfortunately, due to computer server malfunction at WTCCC, access to the relevant data was not possible. Therefore it was decided that alternative strategies should be used to select SNPs for genotyping in the cohort.

In the absence of known SNPs that are correlated with disease severity (aside from *NOD2* SNPs that correlate with stricturing ileal disease[85]), it was decided that the top susceptibility SNPs from a CD meta-analysis[117] should be genotyped in the Dundee cohort. CD and UC patients as well as controls were genotyped and examined for associations with disease severity to enable calculation of the odds ratios for susceptibility in the Dundee cohort. In addition to the *NOD2* 1007fs mutation (rs2066847), the two other common *NOD2* CD SNPs not in the meta-analysis were also genotyped. A list of the SNPs genotyped is given in Table 4-3.

SNP	Chr position	Meta-analysis OR Barrett *et al.*	Gene of interest
rs2476601	1p13.2	1.31	PTPN22
rs11209026	1p31.3	2.50	IL23R
rs2274910	1q23.3	1.14	ITLN1
rs9286879	1q24.3	1.19	?
rs11584383	1q32.1	1.18	?
rs2241880	2q37.1	1.28	ATG16L1
rs3197999	3p21.31	1.20	MST1
rs4613763	5p13.1	1.32	PTGER4
rs2188962	5q31.1	1.25	?
rs13361189	5q33.1	1.33	IRGM
rs10045431	5q33.3	1.11	IL12B
rs6908425	6p22.3	1.21	CDKAL1
rs7746082	6q21	1.17	?
rs2301436	6q27	1.21	CCR6
rs1456893	7p12.2	1.20	?
rs1551398	8q24.13	1.08	?
rs10758669	9p24.1	1.12	JAK2
rs4263839	9q32	1.22	TNFSF15
rs10995271	10q21.2	1.25	ZNF365
rs11190140	10q24.2	1.20	NKX2-3
rs7927894	11q13.5	1.16	C11ORF30
rs11175593	12q12	1.54	LRRK2/MUC19
rs3764147	13q14.11	1.25	?
rs2066844	16q12.1	Not studied	NOD2 R702W
rs2066845	16q12.1	Not studied	NOD2 G908R
rs2066847	16q12.1	3.99	NOD2 1007fs
rs2872507	17q12	1.12	ORMDL3
rs744166	17q21.2	1.18	STAT3
rs2542151	18p11.21	1.35	PTPN2
rs1736135	21q21.1	1.18	?
rs762421	21q22.3	1.13	ICOSLG

Table 4-3 Selected SNPs for genotyping ?=gene not known

4.2.3 Genotyping methods

Genotyping of the SNPs was completed on the Sequenom® platform at the Genomics Core of the University of California, San Francisco, as detailed in Chapter 2. DNA from 366 CD, 261 UC and 539 controls were genotyped.

4.2.3.1 Quality control

The controls were examined for Hardy-Weinberg equilibrium (table 1-4). The SNP rs3764147 was discounted because of a lack of Hardy-Weinberg equilibrium (p-value $3x10^{-4}$). A further SNP, rs4263839, had a poor rate of successful genotyping (62.2% in controls) and was also discounted.

SNP	Gene of interest ?=not known	HW p-value	%Geno	MAF	Maj:Min Alleles
rs2476601	PTPN22	0.4282	100	0.087	G:A
rs11209026	IL23R	0.6825	99.6	0.075	G:A
rs2274910	ITLN1	0.2564	88.3	0.267	C:T
rs9286879	?	0.6135	99.8	0.264	A:G
rs11584383	?	0.8704	99.6	0.313	T:C
rs2241880	ATG16L1	0.5327	99.1	0.492	C:T
rs3197999	MST1	0.7265	97.8	0.283	C:T
rs4613763	PTGER4	1	99.8	0.133	T:C
rs2188962	?	0.4915	98.5	0.475	C:T
rs13361189	IRGM	0.6694	99.3	0.065	T:C
rs10045431	IL12B	0.4013	80.3	0.328	C:A
rs6908425	CDKAL1	0.0061	99.1	0.216	C:T
rs7746082	?	1	100	0.284	G:C
rs2301436	CCR6	0.8572	92	0.465	G:A
rs1456893	?	0.1236	98.3	0.327	A:G
rs1551398	?	0.0839	99.8	0.372	T:C
rs10758669	JAK2	0.7645	99.1	0.358	A:C
rs4263839	TNFSF15	0.681	62.2	0.314	G:A
rs17582416	?	0.5913	99.3	0.343	T:G
rs10995271	ZNF365	0.129	99.4	0.372	G:C
rs11190140	NKX2-3	0.4776	98.9	0.476	C:T
rs7927894	C11ORF30	0.1188	96.5	0.391	C:T
rs11175593	LRRK2/MUC19	1	100	0.019	C:T
rs3764147	?	3.00E-04	19.9	0.182	A:G
rs2066844	NOD2 702	0.9776	100	0.036	C:T
rs2066845	NOD2 908	1	100	0.007	G:C
rs2066847	NOD2 1007	1	100	0.019	Ins C
rs2872507	ORMDL3	0.7651	97	0.482	G:A
rs744166	STAT3	0.3781	98.9	0.46	T:C
rs2542151	PTPN2	1	98.7	0.181	T:G
rs1736135	?	0.2852	99.3	0.447	T:C
rs762421	ICOSLG	0.7548	99.3	0.393	A:G

Table 4-4 Quality control on genotyped SNPs, controls only shown

4.3 Results

4.3.1 Severity score

Of the 366 CD patients in the Dundee cohort, 249 (68.3%) had complete sets of data available for the first 5 years after diagnosis. These patients were used for analysis and validation of the severity score.

4.3.1.1 Score distribution

The graph showing the numbers of patients with each score is shown in Figure 4-1. The Kolmogorov-Smirnow test confirmed that the data approximated a normal distribution (p>0.10). The mean severity score was 6.3 (95% CI 5.90-6.64). A score of $\leq$3 encompassed the least severe 20.9% of the population, whereas a score $\geq$9 identified the most severe 20.5% of the population.

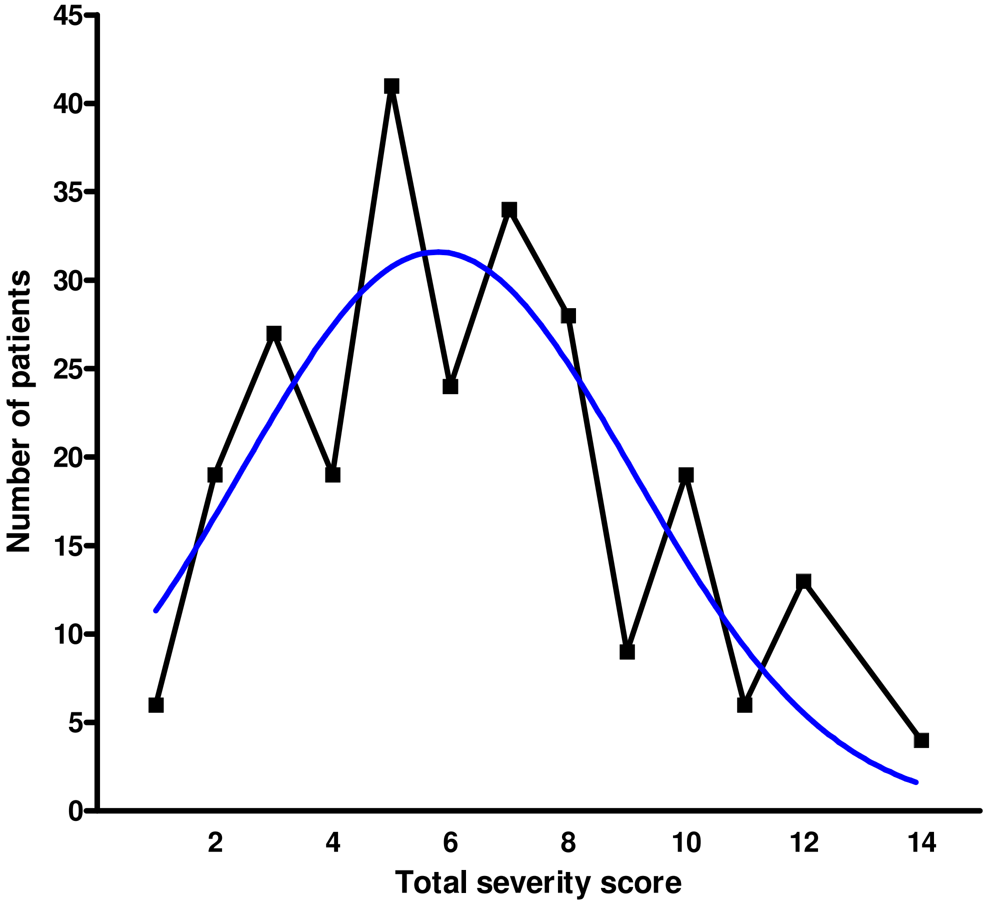


Figure 4-1 Distribution of severity scores, —— line of best fit

4.3.1.2 Sub-score correlation with total score

As the severity score used all important clinical criteria, there were no independent disease parameters with which to compare the scores with long term outcome. Therefore subsets of the score were examined for the best correlation with the total score.

The Pearson correlation was used to examine how individual aspects of the score correlated with the total score, to determine which combination of the criteria had the strongest correlation with the total score. As shown in Table 4-5, the single parameter with the highest correlation with total score was D, the number of hospitalizations. Similarly the 2-parameter combination with the strongest correlation was B+D (medical/surgical management + number of hospitalizations) and the 3-parameter combination was A+B+C (disease extent/behaviour + medical/surgical management + nutritional status).

Parameter	Pearson r	R squared
A Disease extent/behaviour	0.624	0.390
B Medical/surgical management	0.688	0.473
C Nutritional status	0.563	0.317
D Hospitalizations	0.734	0.538
2 Parameters combined		
A+B	0.860	0.739
B+C	0.821	0.674
C+D	0.816	0.665
A+C	0.797	0.635
A+D	0.816	0.667
B+D	0.870	0.756
3 Parameters combined		
A+B+C	0.950	0.903
B+C+D	0.934	0.873
A+C+D	0.904	0.816
A+B+D	0.943	0.889

Table 4-5 Pearson's test for correlation results All p-values for correlation were <0.0001

4.3.1.3 *Long term association*

Based on the two parts of the score that correlated best with the total score: medical/surgical management and hospitalizations, an abbreviated severity score was calculated (total score possible=8) and the patients divided according to their score (lower abbreviated score ≤4, higher abbreviated score >4). A Kaplan-Meier curve was calculated with time to disease progression for each of these groups, as shown in Figure 4-2. The median times to disease progression were 21.2 years and 9.1 years for the lower score and the higher abbreviated score respectively (log rank test p=0.02).

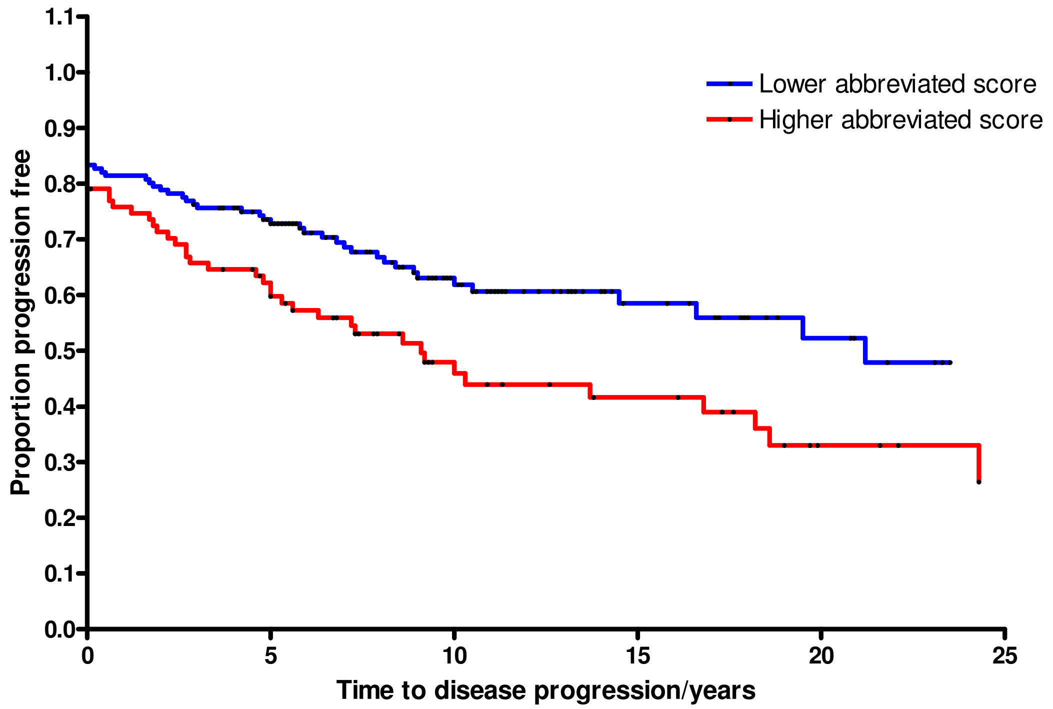


Figure 4-2 Kaplan-Meier: time to disease progression according to abbreviated severity score

4.3.2 Genotyping correlation with severity score

4.3.2.1 *Scoring: total SNP score*

For each patient, all risk alleles were scored (risk allele=1, non-risk allele=0) and summated, and then divided by the total number of alleles genotyped successfully. This calculation gave the percentage of risk alleles present. The scatter plot of this analysis against the severity score is shown in Figure 4-3, which has the line of best fit. A statistically significant correlation was not found between the severity score and the SNP score (two-tailed Pearson's test of correlation, $p = 0.078$, $R = 0.1106$, $R^2 = 0.012$).

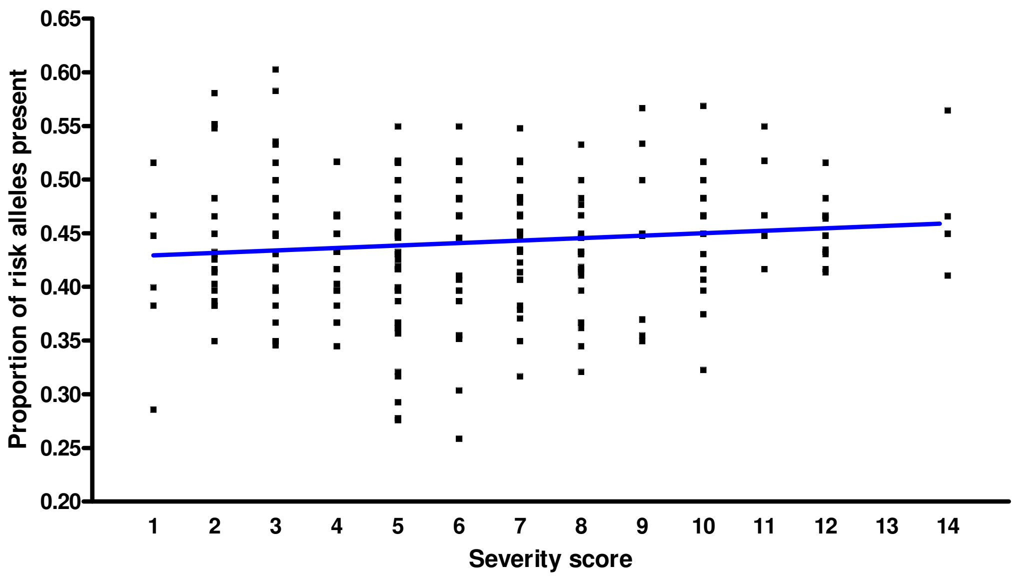


Figure 4-3 Scatter plot of severity score against SNP score. —— line of best fit

4.3.2.2 Scoring: weighted total SNP score

As different SNPs confer different ORs for disease susceptibility, the data was also analysed adding a weighting for the ORs for disease susceptibility from the GWAS meta-analysis.[117] For each SNP, a risk allele homozygote had a score of 2xOR, a risk allele heterozygote had a score of 1xOR and a low risk homozygote had a score of 0. This score was summated and divided by the total possible score if all the successfully genotyped SNPs had been risk allele homozygous. The scatter plot is shown in Figure 4-4. There was no correlation between the weighted SNP and severity scores (Pearson's test of correlation $p = 0.22$).

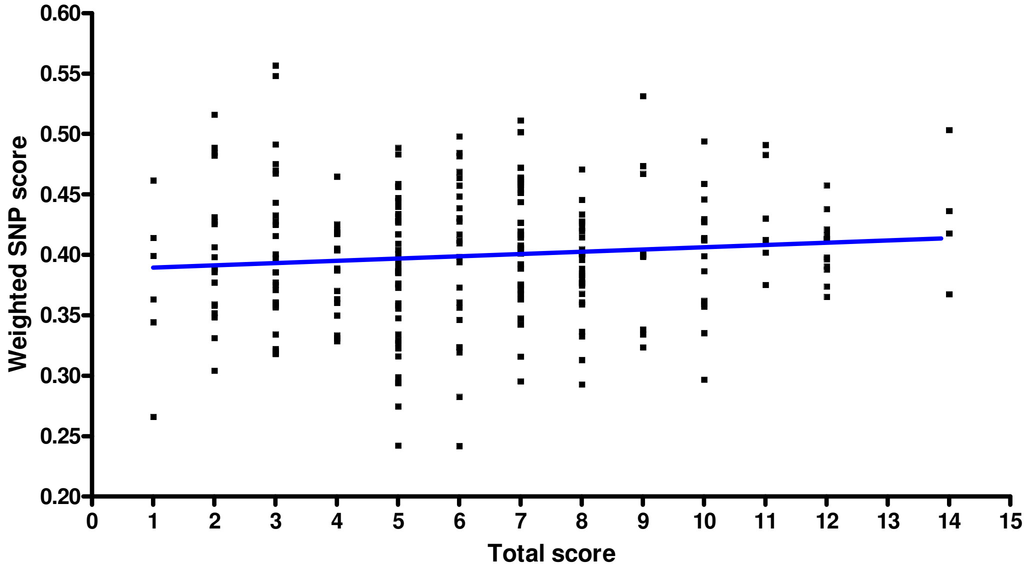


Figure 4-4 Scatter plot of severity score against weighted SNP score, ——— line of best fit

4.3.3 Univariate analysis of factors correlating with more severe disease

The aim of this investigation was to explore whether factors present at diagnosis could be used to predict risk of severe disease in the future and from the model decide at diagnosis the patients most suitable for 'top-down' therapy.

As the mean severity score was 6.3, the 249 patients were subdivided according to severity score: lower score group (score ≤6) and higher score group (score >6). Various factors obtained at diagnosis were each examined for association with more severe disease using the chi-squared test. The list of factors and the results are given in Table 4-6. When corrected for multiple testing with a Bonferroni correction, a significant p-value was <0.005. For smoking at diagnosis and Montreal age group at diagnosis, the p-value is for a chi-squared test for trend.

The genotyping data for each SNP was also analysed using the chi-squared test. For simplicity, SNPs were assumed to behave in a dominant fashion. When corrected for multiple testing with a Bonferroni correction, a significant p-value was <0.002. The results are shown in Table 4-7.

The data demonstrates that A1 age group at diagnosis (i.e.<17 years old at diagnosis) and ileal or Upper GI location at diagnosis had a statistically significant correlation with more severe disease.

In addition, SNP data was analysed according to Montreal behaviour at 5 years. When corrected for multiple testing, none of the SNPs had a statistically significant correlation with B2 or B3 behaviour at 5 years (data not shown).

Factor	OR	95% CI	Chi squared p val
Steroid treatment at diagnosis	1.12	0.677-2.12	0.50
First degree relative with IBD	0.952	0.470-1.94	1.0
Resection at diagnosis	1.53	0.732-3.19	0.300
Sex	0.920	0.554-1.53	0.797
Smoking at diagnosis (smoker, ex-smoker, non smoker)			0.405
Perianal disease at diagnosis	2.23	0.978-5.10	0.065
Montreal age group at diagnosis (A1, A2, A3)			0.0004
A2 compared with A1	0.226	0.072-0.712	
A3 compared with A1	0.131	0.040-0.427	
Ileal location at diagnosis (includes those with ileocolonic disease)	2.20	1.32-3.68	0.0025
Colonic location at diagnosis (includes those with ileocolonic disease)			0.168
Upper GI location at diagnosis	5.30	1.9-14.7	0.0008

Table 4-6 Chi-squared test of clinical factors present at diagnosis and more severe disease

SNP	Chr position	Gene of interest ?=Intergenic area	Overall MAF	OR	CI	p-value
rs2476601	1p13.2	PTPN22	0.094	0.83	0.051 - 13.42	1
rs11209026	1p31.3	IL23R	0.025	0.824	0.233 – 2.92	0.759
rs2274910	1q23.3	ITLN1	0.25	0.461	0.107 - 1.97	0.306
rs9286879	1q24.3	?	0.25	1.77	1.070 - 2.94	0.03
rs11584383	1q32.1	?	0.286	1.51	0.608 - 3.73	0.502
rs2241880	2q37.1	ATG16L1	0.429	1.01	0.517 - 1.96	1
rs3197999	3p21.31	MST1	0.298	1.09	0.648 - 1.82	0.793
rs4613763	5p13.1	PTGER4	0.168	1.27	0.746 - 2.16	0.416
rs2188962	5q31.1	?	0.482	0.934	0.522 - 1.67	0.882
rs13361189	5q33.1	IRGM	0.093	2.09	1.07 - 4.09	0.042
rs10045431	5q33.3	IL12B	0.309	1.36	0.369 - 5.00	0.751
rs6908425	6p22.3	CDKAL1	0.183	1.67	0.409 - 6.86	0.519
rs7746082	6q21	?	0.316	1.20	0.728 - 1.99	0.523
rs2301436	6q27	CCR6	0.47	1.38	0.736 - 2.57	0.346
rs1456893	7p12.2	?	0.317	0.741	0.314 - 1.75	0.517
rs1551398	8q24.13	?	0.351	0.980	0.421 - 2.28	1
rs10758669	9p24.1	JAK2	0.372	1.36	0.806 - 2.31	0.287
rs17582416	10p11.12	?	0.374	1.54	0.921 - 2.57	0.12
rs10995271	10q21.2	ZNF365	0.427	0.949	0.560 - 1.61	0.893
rs11190140	10q24.2	NKX2-3	0.475	0.758	0.439 - 1.31	0.332
rs7927894	11q13.5	C11ORF30	0.426	1.18	0.700 - 2.00	0.595
rs11175593	12q12	LRRK2/MUC19	0.008	1.21	0.075 - 19.5	1
rs2066844	16q12.1	NOD2 R702W	0.044	0.912	0.328 - 2.54	1
rs2066845	16q12.1	NOD2 G908R	0.012	6.25	0.719 - 54.3	0.095
rs2066847	16q12.1	NOD2 1007fs	0.046	1.05	0.477 - 2.31	1
rs2872507	17q12	ORMDL3	0.468	1.24	0.695 - 2.21	0.559
rs744166	17q21.2	STAT3	0.444	0.881	0.469 - 1.66	0.748
rs2542151	18p11.21	PTPN2	0.186	1.03	0.604 - 1.74	1
rs1736135	21q21.1	?	0.372	1.37	0.689 - 2.71	0.396
rs762421	21q22.3	ICOSLG	0.436	0.608	0.350 - 1.06	0.092

Table 4-7 Chi-squared test: SNP genotype with more severe disease

4.3.4 Correlation with Beaugerie severity score

The Beaugerie definition of disabling CD was applied to the 249 patients with full data available. Of these, 218 patients (87.5%) fitted the criteria for disabling disease. To compare the two scoring systems, the sensitivity and specificity of each level of the novel score was calculated (using GraphPad) for its ability to predict whether a patient fitted the Beaugerie disabling disease category or not, and is shown in Table 4-8. This was plotted on a Receiver Operating Characteristic (ROC) curve (Figure

4-5) in order to define the level of the novel score that fitted best with the Beaugerie disabling disease category.

Novel Score Cutoff	Sensitivity	95% CI	Specificity	95% CI	Likelihood ratio
< 1.5	0.188	0.072 - 0.364	1	0.983 - 1.00	
< 2.5	0.656	0.468 - 0.814	0.982	0.954 - 0.995	35.6
< 3.5	0.875	0.710 - 0.965	0.889	0.840 - 0.928	7.91
< 4.5	0.969	0.838 - 0.999	0.816	0.758 - 0.865	5.26
< 5.5	0.969	0.838 - 0.999	0.627	0.559 - 0.691	2.6
< 6.5	0.969	0.838 - 0.999	0.516	0.448 - 0.584	2
< 7.5	1	0.891 - 1.00	0.364	0.300 - 0.432	1.57
< 8.5	1	0.891 - 1.00	0.235	0.180 - 0.297	1.31
< 9.5	1	0.891 - 1.00	0.194	0.143 - 0.253	1.24
< 10.5	1	0.891 - 1.00	0.106	0.068 - 0.155	1.12
< 11.5	1	0.891 - 1.00	0.078	0.046 - 0.123	1.09
< 13.0	1	0.891 - 1.00	0.018	0.005 - 0.047	1.02

Table 4-8 Sensitivity and specificity for the novel score to predict Beaugerie disabling disease

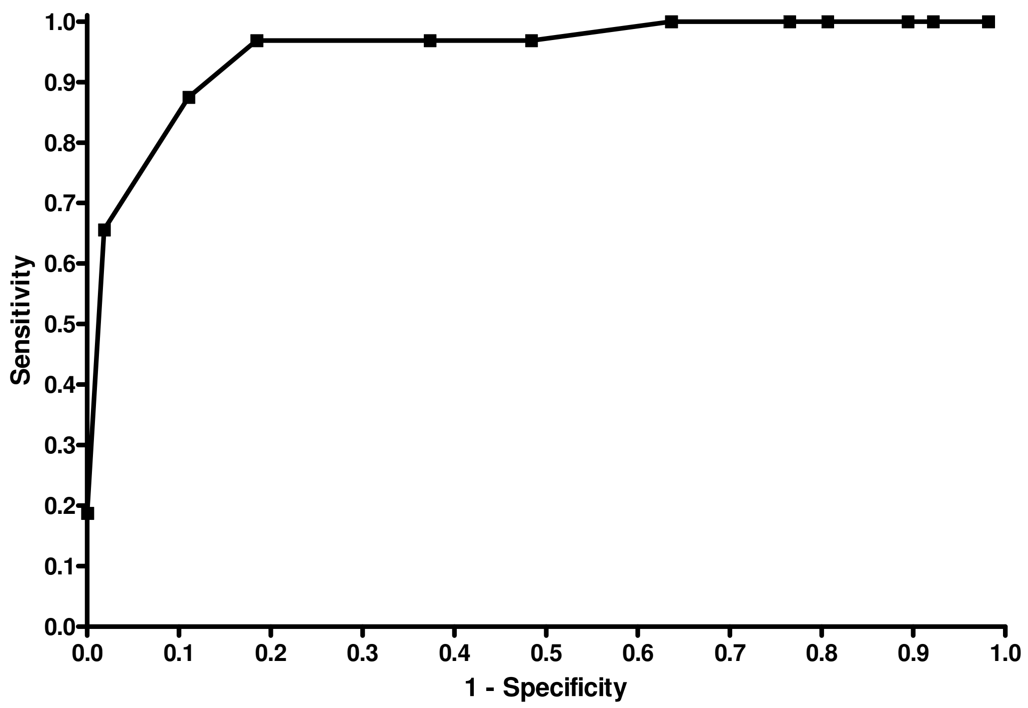


Figure 4-5 ROC curve for Beaugerie severity score compared with the novel severity score

The area under the curve was high at 0.950 (0.912-0.989). A cut-off of >3 in the novel severity score gave the best balance of sensitivity and specificity (sensitivity 0.875, specificity 0.8894) and therefore fitted best with Beaugerie disabling disease.

4.3.5 Multivariate analysis of factors correlating with more severe disease

Factors with a univariate p-value of <0.1 for correlation with more severe disease (perianal disease, age group at diagnosis, ileal and upper GI locations) along with similarly significant SNPs (rs9286879, rs13361189, rs762421, rs17582416 and rs2066845) were tested in a logistic regression model with backward elimination using JMP 8.0.2. As before, SNPs were assumed to behave in a dominant fashion. Independent factors retaining significance (p<0.05) are shown in Table 4-9, along with the OR, positive predictive value (PPV) and negative predictive value (NPV). Age <17 at diagnosis and UGI disease at diagnosis conferred particularly high OR for more severe disease (5.50 and 6.16 respectively).

Factor	OR	PPV	NPV	p-value
Age <17 at Dx	5.50	0.800	0.424	0.002
Ileal at Dx	2.19	0.540	0.348	0.005
Upper GI at Dx	6.16	0.792	0.418	0.0003
rs13361189	2.17	0.605	0.422	0.035
rs9286879	1.99	0.531	0.390	0.014

Table 4-9 Independent factors retaining significance on logistic regression

From this logistic regression, a model was formulated to calculate the probability, from these 5 parameters, of subsequently developing severe disease. (Table 4-10).

Probability of severe disease = 1/(1+Exp[Z]) Z is calculated with the following formula: Z = 1.362
– 1.704 (if <17 at Dx)
– 0.786 (if ileal disease at Dx)
– 1.818 (if UGI disease at Dx)
– 0.776 (if 1-2 rs13361189C alleles)
– 0.688 (if 1-2 rs9286879G alleles)

Table 4-10 Calculating the probability of predicting more severe disease

The ROC curve for this analysis is shown in Figure 4-6, and has an area under the curve of 0.725 signifying a reasonable correlation between the calculated probability of severe disease and the actual severe disease status (more severe vs. less severe). Assuming equal equivalence is given to a false positive and a false negative, the cut off in probability is 0.528, as shown by the yellow line in Figure 4-6, giving a sensitivity of 0.584 and a specificity of 0.75 for the prediction of a severity score of >6 at 5 years.

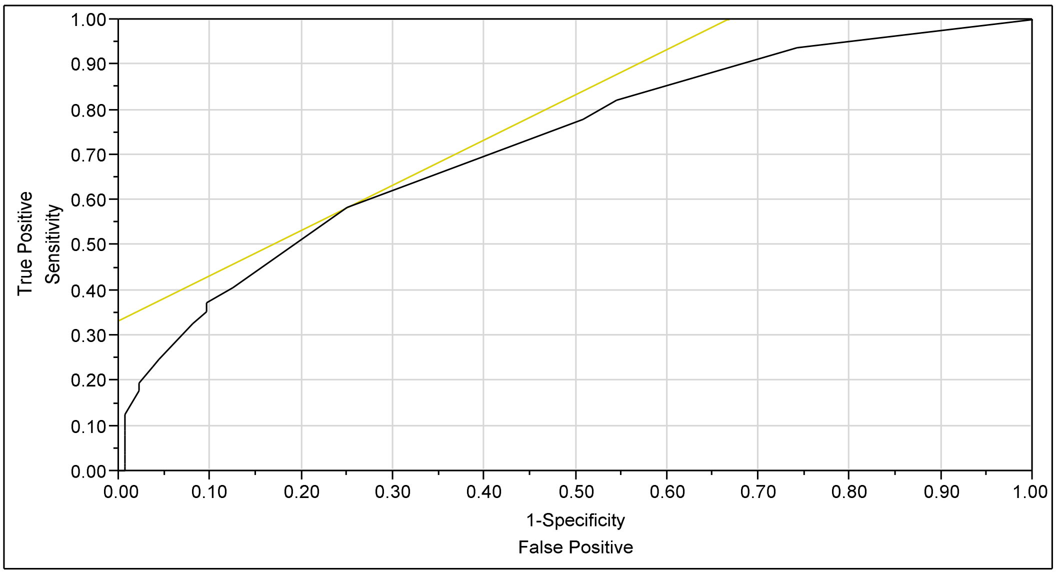


Figure 4-6 ROC curve for the logistic regression model

4.3.6 Genotyping: case control analysis

The results of a case-control analysis of the genotyped SNPs are shown in Table 4-11. As each of these genes had previously shown an association with CD, a significant p-value was <0.05.

SNP	Gene of interest ?=intergenic area	Minor allele	Control MAF	IBD MAF	p-value	CD MAF	p-value	UC MAF	p-value
rs2476601	PTPN22	A	0.087	0.095	0.4517	0.094	0.5755	0.098	0.4674
rs11209026	IL23R	A	0.075	0.032	1.40E-06	0.025	3.97E-06	0.04	0.008
rs2274910	ITLN1	T	0.267	0.252	0.4254	0.25	0.4414	0.264	0.914
rs9286879	?	G	0.264	0.247	0.3519	0.25	0.5049	0.239	0.2923
rs11584383	?	C	0.313	0.278	0.0613	0.286	0.2106	0.276	0.1285
rs2241880	ATG16L1	T	0.492	0.439	0.0088	0.429	0.008	0.462	0.2488
rs3197999	MST1	T	0.283	0.291	0.6641	0.298	0.5039	0.284	0.954
rs4613763	PTGER4	C	0.133	0.148	0.3187	0.168	0.0416	0.13	0.8627
rs2188962	?	T	0.475	0.476	0.9879	0.482	0.7783	0.469	0.82
rs13361189	IRGM	C	0.065	0.085	0.0596	0.093	0.027	0.08	0.2481
rs10045431	IL12B	A	0.328	0.288	0.0649	0.309	0.4611	0.259	0.0164
rs6908425	CDKAL1	T	0.216	0.193	0.1556	0.183	0.0885	0.212	0.8332
rs7746082	?	C	0.284	0.305	0.266	0.316	0.1491	0.287	0.8887
rs2301436	CCR6	A	0.465	0.46	0.8249	0.47	0.8419	0.447	0.5579
rs1456893	?	G	0.327	0.331	0.8183	0.317	0.6641	0.352	0.3073
rs1551398	?	C	0.372	0.358	0.4562	0.351	0.3601	0.362	0.694
rs10758669	JAK2	C	0.358	0.378	0.3149	0.372	0.5403	0.39	0.2242
rs17582416	?	G	0.343	0.357	0.5005	0.374	0.1882	0.351	0.7707
rs10995271	ZNF365	C	0.372	0.41	0.0581	0.427	0.0177	0.391	0.4612
rs11190140	NKX2-3	T	0.476	0.499	0.2653	0.475	0.9629	0.517	0.127
rs7927894	C11ORF30	T	0.391	0.416	0.2133	0.426	0.1443	0.397	0.821
rs11175593	LRRK2/ MUC19	T	0.019	0.014	0.3617	0.008	0.0678	0.015	0.6379
rs2066844	NOD2 R702W	T	0.036	0.033	0.7075	0.044	0.4269	0.022	0.1349
rs2066845	NOD2 G908R	C	0.007	0.01	0.5785	0.012	0.295	0.008	0.963
rs2066847	NOD2 1007fs	C	0.019	0.029	0.1105	0.046	7.00E-04	0.01	0.1723
rs2872507	ORMDL3	A	0.482	0.526	0.0341	0.532	0.0373	0.52	0.1616
rs744166	STAT3	C	0.46	0.427	0.1031	0.444	0.4912	0.393	0.0107
rs2542151	PTPN2	G	0.181	0.2	0.2492	0.186	0.802	0.215	0.1126
rs1736135	?	G	0.447	0.383	0.0014	0.372	0.0013	0.406	0.1185
rs762421	ICOSLG	G	0.393	0.417	0.2326	0.436	0.0703	0.402	0.7236

Table 4-11 Case-control SNP analysis

4.3.6.1 Crohn's disease

The SNPs at or near *IL23R, ATG16L1, IRGM, PTGER4, ZNF365, NOD2* (1007fs), *ORMDL3* and rs1736135 (intergenic area) had a statistically significant association with CD susceptibility compared with controls. The odds ratios (ORs) are shown in Table 4-12. The *IL23R* wild-type allele of rs11209026 conferred the highest risk (OR=3.21), whereas the mutant allele conferred protection (OR=0.31); *NOD2* 1007fs mutations and the autophagy genes, *ATG16L1* and *IRGM,* also were associated with CD susceptibility (ORs 2.57, 1.29 and 1.48 respectively).

4.3.6.2 Ulcerative colitis

The SNPs representing *IL23R, IL12* and *STAT3* had a statistically significant association with UC susceptibility compared with controls. The ORs are given in Table 4-13. Three of these SNPs are in the IL23 pathway (*IL23R*, *IL12B* and *STAT3*) although other IL23 pathway SNPs (*JAK2* and *CCR6)* were not associated with UC susceptibility. *IL12B* is also known as the p40 subunit of both *IL12* (with p35) and *IL23* (with p19).

SNP	Gene of interest	CD OR	95% CI
rs11209026	IL23R	0.31	0.18-0.52
rs2241880	ATG16L1	1.29	1.07-1.56
rs4613763	PTGER4	1.31	1.01-1.70
rs13361189	IRGM	1.48	1.04-2.10
rs10995271	ZNF365	1.26	1.04-1.53
rs2066847	NOD2 1007fs	2.57	1.47-4.50
rs2872507	ORMDL3	1.23	1.01-1.48
rs1736135	Not known	1.37	1.13-1.66

Table 4-12 Odds ratios of statistically significant SNPs in Crohn's disease

SNP	Gene of interest	UC OR	95% CI
rs10045431	IL12B	1.40	1.06-1.83
rs744166	STAT3	1.32	1.07-1.63
rs11209026	IL23R	0.52	0.32-0.85

Table 4-13 Odds ratios of statistically significant SNPs in Ulcerative colitis

4.3.6.3 *Percentage of risk alleles present: case/control comparison*

In order to examine whether CD patients had more risk alleles than controls, the two populations were compared. For each patient, all risk alleles were scored (risk allele=1, non-risk allele=0) and summated, and then divided by the total number of alleles genotyped successfully. This calculation gave the percentage of risk alleles present. The 32 meta-analysis SNPs and the three *NOD2* risk alleles were included in the calculation. The risk allele proportions were compared between CD cases and controls (Figure 4-7). There was a statistically significant difference in the means (CD mean=0.442, 95% CI 0.434-0.449; control mean=0.422, 95% CI 0.417-0.427).

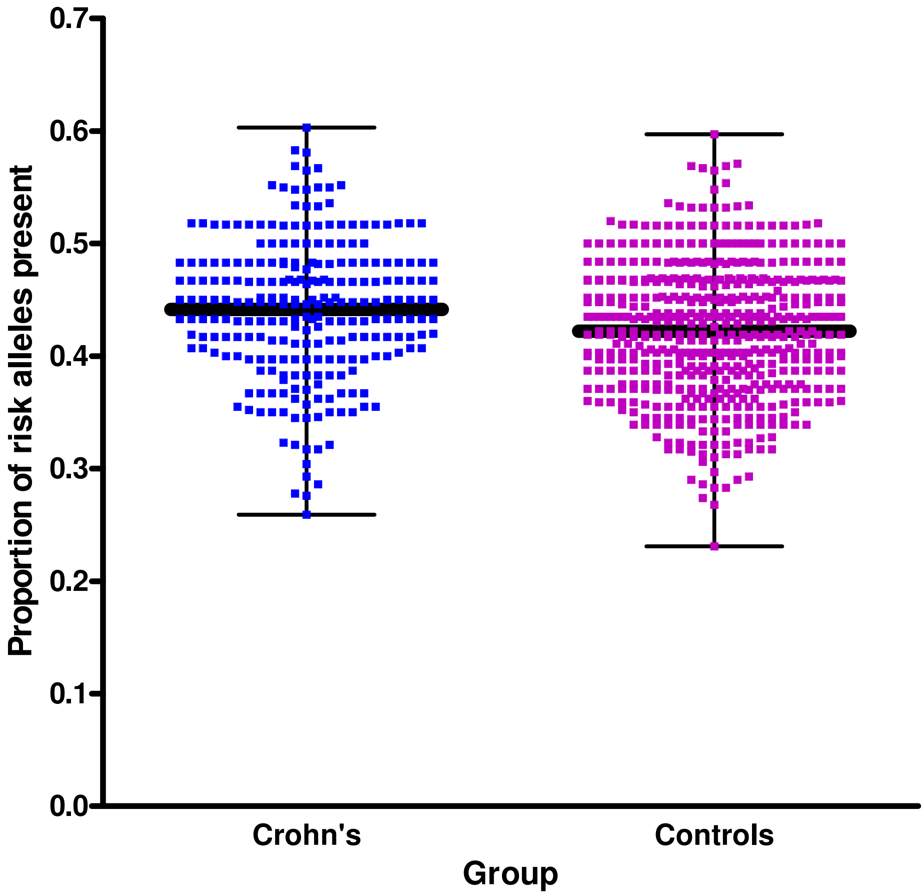


Figure 4-7 Scatter dot plot comparing proportion of risk alleles present in CD and controls; bars represent the upper and lower data points, and thick central line represents mean.

4.4 Discussion

4.4.1 Severity score

The severity score was normally distributed amongst the Dundee CD cohort studied. When defined according to the Beaugerie criteria, a slightly higher proportion of these patients were classified as 'disabling' than in other populations (87.5% vs. 85.2%[234] and 60%[235]), suggesting that Dundee CD patients have a more severe phenotype.

It is difficult to correlate a score that is the summation of the first 5 years of the disease course with long term outcomes, as the parameters which measure disease severity have been used already. To account for this, correlations of subsets of the score with the total score were as examined. The best subset (medical/surgical management and hospitalizations) did not include disease behaviour, and the rates of disease progression were compared between patients with lower and higher abbreviated scores. The statistically significant difference between these 2 groups, strongly suggests that the total score was correlated with a long-term, more severe phenotype.

There are ongoing international efforts to define a Crohn's Disease Digestive Damage Score (the Lémann score).[236] This will be a much more anatomically based score, defining amount of bowel affected by CD with weighting for different locations affected. It will not account for the other aspects of disease severity which are utilised in the severity score detailed in this chapter, for example nutritional status and socio-economic impact:

4.4.2 Severity score correlation with genotyping

There was no correlation between the severity score and proportion of risk alleles present whether the SNPs were weighted or not. With only 32 alleles being examined, this was not surprising, especially since the SNPs are disease susceptibility rather than severity SNPs. It would have been better to have used the SNPs differentiating between B2 and B3 disease in the WTCCC, but these were not available. It would be interesting to compare genome-wide polygenic scoring, as previously analysed in schizophrenia[237] against severity score. It is possible that more severe cases would have more risk alleles.

4.4.3 Univariate analysis of severe disease correlation

A score of >6 was selected to differentiate between patients with more and less severe disease. Choosing the score that separated the patients into halves increased the statistical power of any analyses as well as representing a realistic proportion of CD patients who might be deemed appropriate for top-down therapy. However, the cut-off score could be set at whatever level is felt to represent the proportion of patients who would benefit from early aggressive therapy - this could be further delineated in confirmatory studies in other cohorts.

The only factors examined for correlation with disease severity as a binary variable were those that could be obtained at diagnosis, in keeping with the ultimate aim: to be able to select patients at diagnosis who are at risk of more severe disease. It was surprising that steroid treatment at diagnosis and resection at diagnosis, both clinical decisions, were not correlated with more severe disease. This is in contrast with age group and location at diagnosis which were correlated with disease severity, which are factors that are not based on a clinical decision. The correlation of ileal disease and upper GI disease with disease severity is not surprising considering their strong correlation with disease progression and surgery, as demonstrated in the previous chapter. Perianal disease (p-value 0.052) did not approach statistical significance with Bonferroni correction. The differences in relative contribution of the clinically defined and more objective factors are a reflection of the difficulties of identifying patients with a more severe disease phenotype at diagnosis.

In the Beaugerie study[234], univariate analyses demonstrated that age <40 at diagnosis, pure colonic disease at diagnosis, perianal disease at diagnosis, and steroids at diagnosis corresponded with disabling disease. In contrast, steroids at diagnosis, perianal disease and colonic disease were not correlated with severe disease in the analyses presented here whereas ileal disease and upper GI disease were important in severe disease. The Beaugerie definition encompasses a greater subset of patients, correlating with a novel severity score of ≥3, which may explain the differences in findings.

Although three of the SNPs showed a trend to association with more severe disease, on Bonferroni correction none of the SNPs reached statistical significance.

4.4.4 Multivariate analysis of severe disease correlation

The multivariate analysis confirmed age at diagnosis (especially those <17 at diagnosis) and disease location as important variables as in the univariate analysis. It also confirmed that 2 of the SNPs were correlated with more severe disease: rs13361189 and rs9286878, neither of which had held up to Bonferroni correction on univariate analyses, as detailed in section 4.3.3. The SNP rs13361189 is associated with the *IRGM* gene. The SNP rs9286879 is on chromosome 1q24.3 in an area without identified genes, as yet. Each one of the five factors is an important predictor of more severe disease if present, but lack of them does not definitely mean less severe disease, as shown by the high negative predictive values for each of the factors.

The model allows direct calculation of the probability for more severe disease. At what level the probability cut-off should be for identifying high risk patients can be debated. In a situation where we did not wish patients to miss out on early aggressive therapy, and just wanted to identify the subset of patients with low risk of more severe disease, the probability of severe disease cut-off could be placed so that the sensitivity was high and the specificity low, for example 0.338 (sensitivity 0.938, specificity 0.257). If we did not wish to unnecessarily give patients early aggressive therapy, the probability cut-off could be placed much higher, for example 0.756 (sensitivity 0.230, specificity 0.963). These cut-offs can be varied depending on clinician preferences as well as finances.

4.4.5 Case-control genotyping

In the case-control study, 8 of the loci were associated with susceptibility to CD (as shown in Table 4-12), and 3 associated with UC (as shown in Table 4-13); the only common risk allele that was significant in both UC and CD was *IL23R*. Indeed, all the positive associations in UC were in the Th17 pathway, confirming its significance in UC pathogenesis, as discussed in Chapter 1.

There were broad similarities in the OR for the Dundee cohort significant CD susceptibility SNPs compared with the GWAS meta-analysis data.[117] The SNPs at or near *IL23R*, *IRGM* and *ORMDL3* had marginally higher OR than in the meta-analysis; *ATG16L1*, *ZNF365* and *PTGER4* had similar OR to the meta-analysis with *NOD2* 1007fs demonstrating a lower OR than the meta-analysis, as shown in Table 4-14.

SNP	Gene of interest	CD OR	Meta-analysis[117] OR
rs11209026	IL23R	0.31	0.40
rs2241880	ATG16L1	1.29	1.28
rs4613763	PTGER4	1.31	1.32
rs13361189	IRGM	1.48	1.33
rs10995271	ZNF365	1.26	1.25
rs2066847	NOD2 1007fs	2.57	3.99
rs2872507	ORMDL3	1.23	1.12
rs1736135	?	1.37	1.18

Table 4-14 CD SNP OR compared with meta-analysis[117] OR

The functional significance of *IL23R* and the Th17 pathway, *NOD2, ATG16L1* and *IRGM* in IBD pathogenesis are in innate immunity and the recognition and destruction of intracellular microbes, as has already been discussed in the Chapter 1.

Of the other loci associated with CD, the SNP rs10995271 lies about 7kbp from the 3' end of the *ZNF365* gene and is in LD with it. Mutations in the *ZNF365* gene are known to be linked with uric acid nephrolithiasis[238], but the function of the gene is poorly understood.

PTGER4 (prostaglandin E receptor 4) on chromosome 5, has 3 exons with a gene structure resembling that of the thromboxane and prostacyclin receptors. It was first associated with CD in a European GWAS in 2007.[70] It is one of a family of prostaglandin E receptors (EP1-4). Prostaglandin E (PGE), a prostanoid, acts on this receptor with cyclooxygenase (COX) enzymes catalysing PGE production. It has long been recognised that non steroidal anti-inflammatory drugs (NSAIDs), which inhibit the COX enzymes, can have a deleterious effect on IBD.[239] In a dextran sodium sulphate (DSS) induced mouse model of colitis using PTGER4$^{-/-}$ mice, 3% DSS was sufficient to produce a severe colitis, which only produced a mild colitis in wild type mice or mice deficient in other prostaglandin receptors. Histology of the colon demonstrated epithelial cell loss and crypt damage, suggesting a role in maintaining mucosal integrity.[240] Therefore the role of PTGER4 appears to be protective in the context of gut inflammation, but further investigation is warranted.

The SNP rs2872507 lies near *ORMDL3* (ORM1-like protein) which is part of the ORM (orsomucoid) protein family, negative regulators of sphingolipid synthesis.[241] Sphingolipids protect cell surfaces by forming part of the plasma membrane lipid layer.[242] They are also involved in cell recognition and signal transmission.[242] Genetic variants regulating its expression have been linked with susceptibility to childhood asthma.[243]

The SNP rs1736135 is in an area where no genes have yet been identified. Ongoing gene mapping may produce further answers in the future as to the functional significance of the area it represents.

Given that this study had less than 400 patients, it is impressive that a quarter of the loci were replicated in the Dundee CD population. It is impossible to tell whether the SNPs not showing association with CD susceptibility in the Dundee cohort were true or false negatives. For a SNP with an allelic frequency >10%, confidently ruling-out an association with CD susceptibility would require 7000-10000 case-control pairs (depending on the allelic frequency of the SNP in question) for an OR of 1.1. As Edinburgh and Dundee are only 60 miles apart, the genetic architecture of cases and controls recruited is likely to be very similar between the two populations. To increase the power of the studies it would be useful to genotype CD patients from the Edinburgh cohort for all of the 32 SNPs. However, as many of the GWAS meta-analysis SNPs had an OR <1.1, even with the combined cohorts the study would remain underpowered.

There was a difference in the means of the SNP risk allele score between CD cases and controls. This has previously been noted in another study.[244] There is evidence that for a large GWAS this sort of approach could be relevant. In schizophrenia, a GWAS was completed on 3322 cases and 3587 controls[237]; individual SNPs were compared in the standard case/control way and the polygenic contribution was also examined. This involved using a filtered proportion of the GWAS SNPs, and using those with p-values of <0.5 to produce a SNP score in a discovery cohort. This SNP score was tested in a further target cohort, and was found to be highly correlated with schizophrenia. This approach takes into account the theory that the genetic association of polygenic diseases, including CD, is explained by large numbers of alleles, each of very small levels of significance. It would be interesting to do this analysis on CD GWAS data.

4.5 Conclusions

The severity score is a novel way of considering disease severity. As a summation of four different variables, it provides a unique way of assessing CD which has not been attempted before. Validation of the long term outcome of the score in predicting more severe disease has been presented. In this chapter, more severe disease has been defined as that which encompasses the most severe half of the patient set, making it more discriminant than the Beaugerie disabling disease definition. However, as it is assessed as a continuous variable, it allows a decision to be made as to where 'more severe' disease should be defined, depending on the study it is used in.

Further studies with larger numbers of SNPs are required examining the hypothesis that more severe CD is correlated with more risk SNPs. This would need to be completed on a genome-wide level on a large number of patients. As genome-wide SNP analysis becomes cheaper in the future, if positive results are gained, it may prove to be a cost-effective method of identifying more severe patients.

The multivariate analysis provides an equation for calculating the probability of severe disease. Undoubtedly this model requires validation in a separate cohort, something that is currently being planned in the Edinburgh cohort. However, if validation is obtained, it provides a potentially useful clinical tool in the treatment decision making process when a patient is first diagnosed. From an extensive examination of the literature, this is the first time that a formal model has been used in a CD cohort for predicting disease severity.

Chapter 5 Germ-line variation in GALNT2 and association with IBD

Summary

Aims: To examine for an association between *GALNT2* and IBD susceptibility in a Scottish cohort, and search for a potentially causative mutation.

Methods: Tagging SNPs across the *GALNT2* gene were genotyped on the Illumina® platform in the Edinburgh cohort of 374 CD, 470 UC patients and 879 controls. WTCCC GWAS genotype data across the *GALNT2* gene was examined for gene-wide significance. Replication, where relevant, was completed on the Dundee cohort of 300 CD, 211 UC and 682 controls using the Taqman® platform. PCR amplification was used to sequence the 16 exons of the gene in patients with specific genotypes.

Results: After initial problems with the genotyping on the Illumina® platform, the SNP rs7536663T demonstrated a statistically significant association with CD susceptibility in the Edinburgh cohort (MAF CD 0.351, controls 0.309, OR 1.38, p-value 0.0008). In addition, a four SNP haplotype containing rs7536663 was associated with CD (p-value 0.0017). No other SNP or haplotype block demonstrated a statistically significant association with IBD, UC or CD; a CD and UC sub phenotypic analysis was also negative. Replication at the rs7536663 locus in the Dundee cohort did not demonstrate a statistically significant association (MAF CD 0.326, controls 0.309, p-value 0.469), although the replication cohort was underpowered to rule out an association. Exonic sequencing across the 16 exons failed to demonstrate any non-synonymous SNPs to be associated with the rs7536663T or four SNP haplotype containing rs7536663.

Conclusion: A larger replication cohort is required to confirm or refute an association of rs7536663T with CD susceptibility.

5.1 Introduction

Following translation of proteins from mRNA in the ribosomes, many proteins require post-translation modification. The addition of sugar moieties to make glycoproteins is a modification particularly characteristic of secreted molecules and cell surface receptors. One of the types of sugar modifications that can occur is O-glycosylation: the attachment of N-acetyl-D-galactosamine (GalNAc) to the hydroxyl group of Ser/Thr side chains on protein substrates, as shown in Figure 5-1. This process occurs after transfer of the protein into the Golgi apparatus, thus is completed at a late stage of protein processing. The addition of these sugar groups gives proteins the capacity to resist proteolysis alongside considerable water holding capacity, enabling them to form components of the extracellular matrix and mucosal secretions.

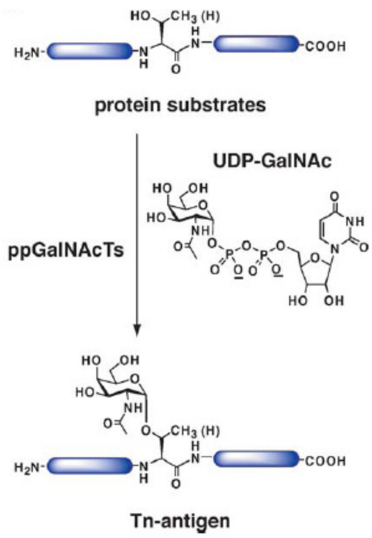


Figure 5-1 O-glycosylation by GALNT enzymes

The initial step in O-glycosylation is catalysed by a family of glycosyltransferases called GALNT (UDP-N-acetyl-α-D-galactosamine polypeptide N-acetylgalactosaminyltransferase).[245] To date 24 human isoforms have been isolated.[246] *GALNT12* SNPs are associated with colorectal cancer susceptibility.[247] Abnormal expression of *GALNT3* is correlated with stage and grade of tumour in a variety of cancers including gastric[248], oesophageal[249], pancreatic[250] and colorectal[251] cancers. SNPs in *GALNT3* contribute to genetic susceptibility to familial hyperphosphataemic tumoral calcinosis (reviewed by Chefetz and Sprecher[252]). The SNP rs4846914, within intron 1 of the *GALNT2* gene, was associated with HDL cholesterol in a GWAS of lipid concentrations and risk of coronary artery disease, conferring an OR of 1.15.[253] It has subsequently been speculated that GALNT2 is important in the O-glycosylation of proteins involved in lipid metabolism.[254] The other GALNT genes have not been associated with other diseases in the literature to date.

5.2 Yeast two-hybrid screen – uncovering protein-protein interactions

The yeast two-hybrid screen is a method of studying protein-protein interactions by analysing the activation of one or more reporter genes in the budding yeast *Saccharomyces cerevisiae*.[255] The protein of interest is cloned into a yeast expression vector and this bait protein used to probe a cDNA library in yeast cells. An interaction causes growth and expansion of the yeast cell population. This method has been used successful to delineate protein interactions and pathways in a variety of conditions including cancer[256], HIV[257] and schizophrenia.[258] Another research group used NOD2 as the bait protein in a yeast two-hybrid screen which suggested that CENTB1[259] and GRIM19[260] were interacting proteins. In the Gastrointestinal laboratory at the University of Edinburgh a yeast two-hybrid screen was also completed by Dr Elaine Nimmo using NOD2 as the bait protein[261], which identified six proteins that appeared to interact with NOD2, as shown in Table 5-1.

Gene	Chr	Function
GALNT2	1q41-42	O-linked glycosylation
TLE1	9q21.32	Transcriptional corepressor - binds to a number of transcription factors
Vimentin	10p12.33	Maintains cell shape in integrity, stabilises cytoskeleton interactions, involved in immune response, controls protein transport
FIS1	7q22.1	Promotes the fragmentation of the mitochondrial network and its perinuclear clustering
HTATIP	11q13	Catalytic subunit of NuA4 histone acetyltransferase complex, involved in transcriptional activation of select genes
PPP2R5E	14q23.2	Belongs to the phosphatase 2A regulatory subunit B family, implicated in the negative control of cell growth and division

Table 5-1 Yeast two-hybrid screen NOD2 interacting proteins

Thus GALNT2 (UDP-N-acetyl-α-D-galactosamine polypeptide N-acetylgalactosaminyltransferase) appeared to be one of the interacting proteins. The *GALNT2* gene is located on chromosome 1q41-42.[262] The coding sequence is 1713bp arranged in 16 exons spanning 100Kbp, with the gene encoding a protein of 571 amino acids (64.7 kDa).[263] The protein structure is thought to consist of a large glycosyl transferase segment lying within the lumen of the Golgi apparatus, a transmembrane anchor and a short cytoplasmic domain.[246]

In the event of an interaction of NOD2 with GALNT2, it follows that mutations in the *GALNT2* gene may predispose to IBD susceptibility. This chapter sets out the investigation of this hypothesis. This hypothesis has plausibility because changes in mucus composition due to defective O-glycosylation could affect barrier protection of the gastrointestinal tract.

5.3 Methods

5.3.1 GALNT2 tagging SNP analysis

Tagging SNPs across the *GALNT2* gene had already been selected by Dr Elaine Nimmo. The 27 SNPs chosen are given in Table 5-2. These SNPs were genotyped on the Illumina® platform at the Wellcome Trust Clinical Research Facility (WTCRF) at the Western General Hospital in the Edinburgh cohort of 374 CD patients, 470 UC patients and 879 controls.

SNP	Alleles	Chr 1 position (B36)
rs12135308	A:T	228253271
rs7515331	T:C	228253467
rs1474925	G:T	228274056
rs910820	A:G	228280942
rs11122300	G:A	228287788
rs12029443	C:G	228293831
rs7536663	A:T	228295422
rs2057233	G:A	228299413
rs868287	G:C	228307737
rs1474839	G:A	228308023
rs1884578	A:G	228333832
rs1973943	C:T	228338686
rs4846908	G:A	228351679
rs1998064	A:C	228353751
rs2281719	A:G	228364282
rs4846919	G:A	228368074
rs1264084/rs606587	A:G	228382812
rs4846928	C:T	228408528
rs7512360	G:A	228422806
rs3811486	T:C	228438684
rs901675	C:T	228456625
rs1471915	C:T	228469247
rs2273967	G:A	228481916
rs16851339	A:T	228483455
rs1043908	T:C	228483917
rs1043909	G:A	228483994
rs7022	G:C	228484287

Table 5-2 GALNT2 SNP selection

In addition, the genotyping across *GALNT2* generated by the WTCCC GWAS[73] in UK CD cases and controls was analysed looking at the p-values for the SNPs across the gene.

5.3.2 PCR optimization for GALNT2 sequencing

The PCR amplifications were optimized using gradient gels to check the working temperature for the PCR and the best quantities of reagents. DNA from a patient (0309UC) was used as the DNA template for amplification. Some of the gradient PCRs required increased amounts of Taq (806/807: failed PCR) or increased magnesium (818/819: non specific product) as shown in Table 5-3, which also shows primer combinations.

Exon	Primer pair	Forward' primer sequence	Reverse primer sequence	PCR temp	
1	804/ 805	AAGACCAAGGCTGAAGAGG	GACAAGACTCGAATTACGACG	61ºC	Didn't work well
1	853/ 854	GATTGGTCGCCCTTTGCTC	GAAAATGCAAGCGCAGGAG	65ºC	Didn't work well
1	853/ 805	GATTGGTCGCCCTTTGCTC	GACAAGACTCGAATTACGACG	65ºC	Betaine 0.1M
2	715/ 716	CTTGCATCTTGACTCCCTAAC	AAGTTTTAACTTGACCACCCAC	61ºC	
3	765/ 766	TTCTGGAGTTAACACTTACTGCAG	GCCAGTCACACAACTCTCTTTC	65ºC	
4-6	806/ 807	CTCTGAAGGGCTAAACAAGC	TTGAGCTAAAGCACTTCTTTATG	62ºC	2xTaq
7	808 /809	TGGAGCAGTTTTGATCTCATC	ACTAATCAGGCTCTCACCAGC	65ºC	
8	810/ 811	GTAGTAGGTAGTAAGGGCCAGTTG	GAGGGCTTCATCATCTCTG	61ºC	
9	812/ 813	AGTCCCCTTCTCTCTCCTCC	CAACATACAGGAATGCACAGAC	65ºC	
10	814/ 815	CACTGTGGGAACATCCAGG	AAACACTTGGGTTCAGACGC	64ºC	
11	816/ 817	TCCTTATCAGATCTCCTTCAGC	TTCACACTCATGAAGATTTGTCC	65ºC	
12-13	818/ 819	CAAATGGAGTGATGCAGACAG	ATTCTCAGCTCCACACCATG	65ºC	2xMgCl2
14	820/ 821	TGTGGTAGGAAGAGGCACG	GTTGTATAGGCATACCAGGCAG	64ºC	
15	822/ 823	TGAATCTAAGCTCCACCCC	CACTTGCATTATGTATCGAGTACC	62ºC	
16	824/ 825	TCCTGAATTCACACGAATCTG	ATTTCAAGAGTTCTCCTCGCTC	60ºC	

Table 5-3 GALNT2 primer sequences and conditions

There were particular problems with the amplification of exon 1, which is discussed in 5.8.1. Once PCRs were optimized, PCRs were completed on samples as discussed in Sections 5.8.2 and 5.8.3.

5.4 Initial Illumina® GALNT2 data analysis

5.4.1 Quality control

GALNT2 data from the Illumina® Goldengate platform were initially subject to quality control. All the SNPs were in Hardy-Weinberg equilibrium in controls. The SNP rs868287 was excluded due to the poor rate of genotyping (86% genotyping). When poorly genotyped DNA (5 samples) was also excluded, it left 371 CD, 470 UC and 879 controls for analysis across the 26 remaining SNPs.

5.4.2 GALNT2 Single SNP analysis

The initial results for the single SNP analysis are shown in Table 5-4. After Bonferroni correction, a significant p-value was taken as <0.0019. As can be seen, rs2281719 was the only significant CD SNP (p-value of 0.001 in IBD, 0.0002 in CD); it showed a non-significant p-value in UC. The MAF was 0.474 in CD and 0.394 in controls.

SNP	Minor allele	Control MAF	IBD MAF	p-value	CD MAF	p-value	UC MAF	p-value
rs12135308	T	0.107	0.111	0.739	0.125	0.2087	0.098	0.4503
rs7515331	C	0.107	0.122	0.173	0.121	0.3355	0.119	0.3759
rs1474925	T	0.213	0.209	0.781	0.207	0.7329	0.205	0.6354
rs910820	G	0.209	0.219	0.501	0.23	0.2521	0.209	0.9949
rs11122300	A	0.189	0.185	0.76	0.174	0.3764	0.195	0.7285
rs12029443	G	0.162	0.168	0.627	0.148	0.3928	0.183	0.1699
rs7536663	T	0.312	0.323	0.494	0.365	0.012	0.292	0.3201
rs2057233	A	0.398	0.383	0.368	0.344	0.0112	0.416	0.3838
rs1474839	A	0.179	0.179	0.951	0.155	0.1323	0.199	0.2111
rs1884578	G	0.269	0.278	0.577	0.319	0.0128	0.247	0.2263
rs1973943	T	0.448	0.418	0.087	0.456	0.7168	0.387	0.0038
rs4846908	A	0.443	0.448	0.758	0.457	0.5104	0.446	0.8779
rs1998064	C	0.219	0.232	0.361	0.263	0.0172	0.209	0.555
rs2281719	G	0.394	0.449	0.001	0.474	0.0002	0.431	0.0661
rs4846919	A	0.151	0.144	0.582	0.151	0.9818	0.14	0.4422
rs1264084/rs606587	G	0.337	0.33	0.699	0.333	0.8585	0.333	0.852
rs4846928	T	0.142	0.146	0.751	0.154	0.4253	0.14	0.8967
rs7512360	A	0.441	0.424	0.34	0.435	0.7867	0.415	0.2125
rs3811486	C	0.116	0.113	0.777	0.113	0.8175	0.113	0.7711
rs901675	T	0.297	0.285	0.46	0.263	0.087	0.3	0.836
rs1471915	T	0.114	0.117	0.726	0.106	0.5789	0.128	0.2644
rs2273967	A	0.253	0.253	0.985	0.25	0.8697	0.255	0.9101
rs16851339	T	0.021	0.016	0.28	0.015	0.3946	0.015	0.3103
rs1043908	C	0.121	0.12	0.979	0.102	0.1893	0.131	0.4426
rs1043909	A	0.188	0.202	0.3	0.193	0.7838	0.212	0.1414
rs7022	C	0.325	0.319	0.708	0.314	0.5915	0.319	0.7575

Table 5-4 GALNT2 Single SNP Analysis

5.4.3 GALNT2 Haplotype analysis

Haplotypes were defined according to solid spine of LD on the control samples. The LD plot is shown in Figure 5-2, and the analysis shown in Table 5-2. The only significant haplotype was the GG in block 8, containing rs2281719 and rs4846919. On its own, rs2281719 had a CD p-value of 0.0002. The p-value improved slightly to 8.9×10^{-5} in the GG haplotype block.

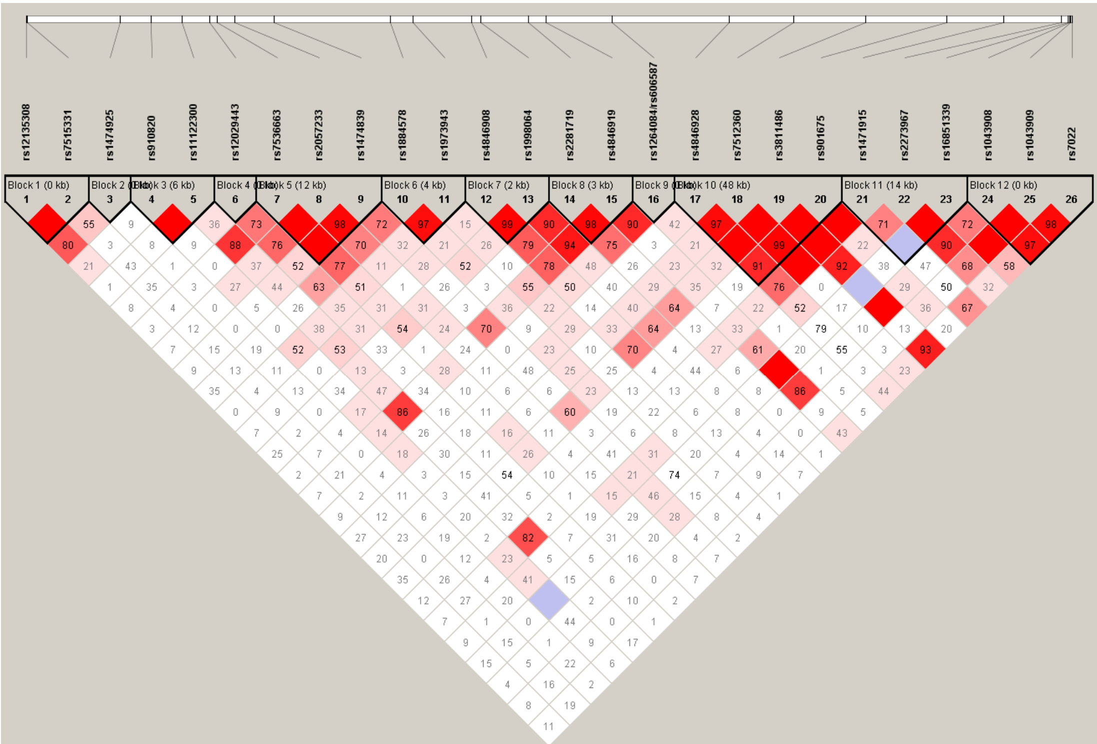


Figure 5-2 GALNT2 LD plot: controls only

Block 1	Controls	IBD	p-value	CD	p-value	UC	p-value
AT	0.785	0.771	0.3095	0.755	0.0963	0.784	0.9275
TT	0.107	0.119	0.2653	0.124	0.2185	0.118	0.3878
AC	0.107	0.11	0.8238	0.121	0.333	0.098	0.4433
Block 2							
G	0.787	0.794	0.6141	0.793	0.7347	0.795	0.6385
T	0.213	0.206	0.6141	0.207	0.7347	0.205	0.6385
Block 3							
AG	0.602	0.597	0.7598	0.597	0.8101	0.597	0.802
GG	0.21	0.219	0.5211	0.23	0.2453	0.209	0.9627
AA	0.189	0.185	0.7712	0.173	0.3526	0.195	0.7181
Block 4							
C	0.838	0.833	0.6747	0.852	0.3945	0.817	0.1731
G	0.162	0.167	0.6747	0.148	0.3945	0.183	0.1731
Block 5							
TGG	0.313	0.324	0.4713	0.365	0.0112	0.29	0.2587
AGG	0.288	0.292	0.8439	0.289	0.986	0.295	0.7879
AAG	0.219	0.205	0.3037	0.192	0.1218	0.217	0.8475
AAA	0.178	0.178	0.9915	0.154	0.1496	0.198	0.2194
Block 6							
AC	0.544	0.562	0.2949	0.523	0.34	0.593	0.0151
GT	0.262	0.278	0.2892	0.316	0.0058	0.247	0.4068
AT	0.19	0.158	0.0142	0.158	0.0571	0.159	0.0427
Block 7							
GA	0.557	0.549	0.646	0.542	0.507	0.554	0.9044
AC	0.22	0.233	0.3429	0.263	0.0203	0.236	0.456
AA	0.223	0.218	0.6842	0.195	0.1129	0.209	0.5388
Block 8							
AG	0.604	0.55	0.0013	0.527	3.00E-04	0.569	0.0794
GG	0.245	0.303	1.00E-04	0.32	8.93E-05	0.29	0.0118
GA	0.15	0.146	0.7406	0.152	0.8696	0.141	0.545
Block 9							
A	0.663	0.667	0.8227	0.667	0.859	0.667	0.8527
G	0.337	0.333	0.8227	0.333	0.859	0.333	0.8527
Block 10							
CGTT	0.293	0.283	0.4674	0.261	0.1043	0.301	0.7292
CGTC	0.268	0.294	0.0946	0.305	0.056	0.284	0.3849
CATC	0.18	0.164	0.2083	0.163	0.2997	0.165	0.3086
TATC	0.139	0.146	0.5539	0.155	0.2849	0.138	0.9616
CACC	0.117	0.113	0.7481	0.113	0.7978	0.112	0.755
Block 11							
CGA	0.643	0.64	0.8454	0.654	0.6142	0.628	0.4591
CAA	0.222	0.226	0.7893	0.224	0.8989	0.228	0.7653
TGA	0.104	0.108	0.7076	0.096	0.5621	0.118	0.2862
CAT	0.021	0.016	0.2479	0.016	0.3934	0.016	0.3309
Block 12							
TGG	0.486	0.478	0.6234	0.49	0.8671	0.467	0.3594
TGC	0.205	0.201	0.7812	0.215	0.5615	0.19	0.3407
TAG	0.187	0.201	0.3025	0.189	0.897	0.212	0.1408
CGC	0.118	0.114	0.7306	0.097	0.1338	0.128	0.4596

Table 5-5 GALNT2 Haplotype analysis

5.4.4 Sub phenotypic analysis - CD

Results of a CD sub phenotypic analysis are shown in Table 5-6 and were negative after Bonferroni correction (p-value <0.0019). A CD sub phenotypic haplotype analysis (Table 5-7 and Table 5-8) revealed that the GG haplotype in block 8 was significant in L2 disease with a p-value of $3x10^{-4}$.

SNP	Minor allele	Control MAF	L1 MAF	p-value	L2 MAF	p-value	L3 MAF	p-value	B1 MAF	p-value	B2 MAF	p-value	B3 MAF	p-value
rs12135308	T	0.107	0.128	0.3444	0.12	0.3289	0.105	0.9402	0.13	0.0563	0.118	0.582	0.115	0.6088
rs7515331	C	0.107	0.123	0.4891	0.129	0.2843	0.092	0.5594	0.116	0.5852	0.085	0.4565	0.182	0.0104
rs1474925	T	0.213	0.214	0.9549	0.191	0.6081	0.169	0.2093	0.204	0.8143	0.204	0.9447	0.224	0.628
rs910820	G	0.209	0.268	0.0464	0.236	0.2994	0.209	0.9993	0.221	0.5371	0.218	0.824	0.292	0.0263
rs11122300	A	0.189	0.167	0.4165	0.176	0.6065	0.189	0.9985	0.16	0.1963	0.191	0.9673	0.188	0.9145
rs12029443	G	0.162	0.152	0.698	0.156	0.7632	0.151	0.7335	0.133	0.3007	0.152	0.678	0.2	0.3198
rs7536663	T	0.312	0.389	0.0227	0.357	0.165	0.356	0.2681	0.343	0.5051	0.361	0.2736	0.402	0.036
rs2057233	A	0.398	0.314	0.0152	0.354	0.1689	0.388	0.8051	0.346	0.0558	0.33	0.1303	0.356	0.2811
rs1474839	A	0.179	0.152	0.3067	0.16	0.4614	0.184	0.8829	0.137	0.074	0.152	0.443	0.209	0.4121
rs1884578	G	0.269	0.33	0.0537	0.312	0.1852	0.322	0.157	0.302	0.3998	0.324	0.2251	0.364	0.0196
rs1973943	T	0.448	0.445	0.9425	0.439	0.7507	0.486	0.3737	0.429	0.326	0.433	0.733	0.54	0.0514
rs4846908	A	0.443	0.465	0.5347	0.457	0.6272	0.467	0.5723	0.468	0.1751	0.438	0.9975	0.417	0.634
rs1998064	C	0.219	0.259	0.1737	0.271	0.0971	0.273	0.1231	0.266	0.0373	0.282	0.1144	0.227	0.8249
rs2281719	G	0.394	0.527	0.0249	0.492	0.0038	0.447	0.1977	0.473	0.0122	0.429	0.5117	0.492	0.0344
rs4846919	A	0.151	0.162	0.6591	0.134	0.5274	0.191	0.1912	0.138	0.4855	0.13	0.5679	0.205	0.083
rs1264084/rs606587	G	0.337	0.353	0.6314	0.324	0.6707	0.327	0.8053	0.333	0.8075	0.339	0.9816	0.321	0.7001
rs4846928	T	0.142	0.17	0.2638	0.15	0.9199	0.125	0.5703	0.155	0.6278	0.145	0.9937	0.154	0.8064
rs7512360	A	0.441	0.441	0.9832	0.449	0.8129	0.39	0.2395	0.45	0.9867	0.409	0.5071	0.439	0.9428
rs3811486	C	0.116	0.112	0.8316	0.102	0.5712	0.14	0.3911	0.121	0.8308	0.098	0.5962	0.106	0.7733
rs901675	T	0.297	0.257	0.214	0.26	0.1797	0.293	0.9336	0.261	0.0905	0.268	0.4833	0.276	0.5595
rs1471915	T	0.114	0.111	0.8965	0.109	0.5739	0.12	0.811	0.095	0.1296	0.116	0.9546	0.142	0.371
rs2273967	A	0.253	0.239	0.6387	0.267	0.4908	0.274	0.5796	0.251	0.9105	0.208	0.3165	0.282	0.4735
rs16851339	T	0.021	0.018	0.8019	0.008	0.3128	0.021	0.9995	0.021	0.6421	0.019	0.9822	0	0.1148
rs1043908	C	0.121	0.098	0.3256	0.102	0.6305	0.097	0.4022	0.091	0.2003	0.111	0.8328	0.125	0.8361
rs1043909	A	0.188	0.205	0.5626	0.168	0.4107	0.185	0.9206	0.189	0.922	0.213	0.541	0.18	0.791
rs7022	C	0.325	0.306	0.5772	0.32	0.9609	0.284	0.3042	0.302	0.4892	0.33	0.9127	0.31	0.7133

Table 5-6 GALNT2 CD sub phenotypic analysis

Haplotype	Control freq	L1 freq	p-value	L2 freq	p-value	L3 freq	p-value
Block 1							
AT	0.785	0.75	0.2285	0.751	0.1241	0.803	0.6174
AC	0.107	0.123	0.4856	0.129	0.2847	0.092	0.5584
TT	0.107	0.127	0.3695	0.12	0.3319	0.105	0.9371
Block 2							
G	0.787	0.786	0.9551	0.809	0.6095	0.83	0.2151
T	0.213	0.214	0.9551	0.191	0.6095	0.17	0.2151
Block 3							
AG	0.602	0.567	0.3208	0.587	0.6396	0.602	0.9964
GG	0.21	0.267	0.0486	0.236	0.2974	0.211	0.9706
AA	0.189	0.166	0.4038	0.176	0.615	0.187	0.965
Block 4							
C	0.838	0.848	0.6992	0.844	0.764	0.849	0.7336
G	0.162	0.152	0.6992	0.156	0.764	0.151	0.7336
Block 5							
TGG	0.313	0.387	0.0234	0.357	0.165	0.359	0.233
AGG	0.289	0.294	0.873	0.287	0.9744	0.252	0.3381
AAG	0.22	0.166	0.066	0.197	0.4092	0.204	0.6597
AAA	0.178	0.152	0.3397	0.159	0.4413	0.184	0.8477
Block 6							
AC	0.544	0.523	0.5609	0.54	0.9602	0.499	0.292
GT	0.261	0.326	0.0404	0.31	0.1148	0.315	0.1511
AT	0.191	0.149	0.1302	0.148	0.1034	0.179	0.701
Block 7							
GA	0.557	0.535	0.5426	0.543	0.6233	0.531	0.5397
AC	0.22	0.259	0.1856	0.271	0.1071	0.273	0.1312
AA	0.223	0.206	0.5478	0.186	0.285	0.196	0.4343
Block 8							
AG	0.604	0.531	0.036	0.509	0.0049	0.552	0.2086
GG	0.244	0.304	0.0508	0.355	3.00E-04	0.257	0.7252
GA	0.15	0.164	0.583	0.135	0.6149	0.19	0.1853
Block 9							
A	0.663	0.647	0.6329	0.676	0.6719	0.673	0.8065
G	0.337	0.353	0.6329	0.324	0.6719	0.327	0.8065
Block 10							
CGTT	0.294	0.256	0.2407	0.255	0.1621	0.296	0.9616
CGTC	0.268	0.304	0.2426	0.297	0.2596	0.314	0.2154
CATC	0.181	0.158	0.4013	0.19	0.6487	0.127	0.0914
TATC	0.139	0.17	0.2037	0.15	0.8022	0.125	0.6324
CACC	0.117	0.111	0.7948	0.104	0.6274	0.138	0.4279
Block 11							
CGA	0.644	0.667	0.5039	0.636	0.8493	0.611	0.45
CAA	0.221	0.204	0.5618	0.246	0.3398	0.248	0.4737
TGA	0.103	0.094	0.6854	0.099	0.5359	0.114	0.7078
CAT	0.021	0.019	0.7884	0.01	0.3308	0.021	0.954
Block 12							
TGG	0.486	0.488	0.9683	0.501	0.7118	0.531	0.286
TGC	0.205	0.209	0.8676	0.228	0.4903	0.187	0.5991
TAG	0.187	0.198	0.6854	0.168	0.4228	0.184	0.926
CGC	0.118	0.094	0.28	0.094	0.4384	0.096	0.4185

Table 5-7 GALNT2 CD sub phenotypic haplotype analysis – disease location

Haplotype	Control freq	B1 freq	p-value	B2 freq	p-value	B3 freq	p-value
Block 1							
AT	0.785	0.755	0.0679	0.795	0.901	0.704	0.0202
AC	0.107	0.116	0.5844	0.087	0.4809	0.181	0.0108
TT	0.107	0.129	0.0594	0.118	0.5827	0.115	0.6291
Block 2							
G	0.787	0.796	0.8153	0.796	0.9456	0.776	0.6281
T	0.213	0.204	0.8153	0.204	0.9456	0.224	0.6281
Block 3							
AG	0.602	0.618	0.6552	0.592	0.8817	0.529	0.1097
GG	0.21	0.222	0.523	0.217	0.8543	0.292	0.0258
AA	0.189	0.16	0.213	0.192	0.9963	0.179	0.7379
Block 4							
C	0.838	0.867	0.3008	0.848	0.678	0.801	0.3266
G	0.162	0.133	0.3008	0.152	0.678	0.199	0.3266
Block 5							
TGG	0.313	0.344	0.5027	0.358	0.305	0.401	0.031
AGG	0.289	0.307	0.2271	0.312	0.572	0.242	0.2717
AAG	0.22	0.212	0.6361	0.179	0.281	0.148	0.043
AAA	0.178	0.137	0.0727	0.152	0.4546	0.209	0.3987
Block 6							
AC	0.544	0.55	0.5826	0.53	0.8149	0.438	0.0208
GT	0.261	0.3	0.2872	0.317	0.1985	0.361	0.0114
AT	0.191	0.149	0.0622	0.142	0.186	0.2	0.8709
Block 7							
GA	0.557	0.532	0.1729	0.562	0.9903	0.58	0.6782
AC	0.22	0.266	0.039	0.281	0.1194	0.228	0.8336
AA	0.223	0.202	0.637	0.156	0.1123	0.192	0.477
Block 8							
AG	0.604	0.529	0.0142	0.571	0.5303	0.513	0.0474
GG	0.244	0.331	0.0013	0.297	0.253	0.28	0.4599
GA	0.15	0.14	0.5275	0.131	0.6169	0.207	0.0636
Block 9							
A	0.663	0.667	0.8083	0.661	0.9816	0.679	0.7001
G	0.337	0.333	0.8083	0.339	0.9816	0.321	0.7001
Block 10							
CGTT	0.294	0.26	0.0906	0.268	0.5182	0.276	0.5962
CGTC	0.268	0.293	0.0987	0.323	0.1757	0.287	0.5388
CATC	0.181	0.168	0.5072	0.16	0.5848	0.178	0.9347
TATC	0.139	0.155	0.5761	0.151	0.7869	0.154	0.7037
CACC	0.117	0.122	0.8259	0.098	0.595	0.104	0.7282
Block 11							
CGA	0.644	0.66	0.4345	0.689	0.3459	0.592	0.261
CAA	0.221	0.222	0.9006	0.175	0.258	0.265	0.2714
TGA	0.103	0.088	0.163	0.102	0.954	0.127	0.4609
CAT	0.021	0.022	0.671	0.02	0.9909	0.001	0.1206
Block 12							
TGG	0.486	0.508	0.6241	0.457	0.5661	0.5	0.7317
TGC	0.205	0.212	0.7684	0.219	0.7668	0.195	0.7375
TAG	0.187	0.184	0.9973	0.202	0.6948	0.18	0.8119
CGC	0.118	0.086	0.1419	0.101	0.6368	0.124	0.7941

Table 5-8 GALNT2 CD sub phenotypic haplotype analysis - disease behaviour

5.4.5 Sub phenotypic analysis – UC

Results of a UC sub phenotypic analysis of GALNT are shown in Table 5-9 and Table 5-10. When corrected for multiple testing there are no significant results.

	Minor allele	Control MAF	E3 MAF	p-value	E2 MAF	p-value	E1 MAF	p-value
rs12135308	T	0.107	0.116	0.6379	0.083	0.1748	0.101	0.8242
rs7515331	C	0.107	0.126	0.2877	0.117	0.6036	0.115	0.7759
rs1474925	T	0.213	0.209	0.8701	0.203	0.6973	0.209	0.9276
rs910820	G	0.209	0.264	0.0234	0.181	0.2227	0.171	0.2847
rs11122300	A	0.189	0.196	0.762	0.201	0.6037	0.182	0.8388
rs12029443	G	0.162	0.203	0.0552	0.169	0.7245	0.169	0.8241
rs7536663	T	0.312	0.279	0.2186	0.305	0.8107	0.288	0.5497
rs2057233	A	0.398	0.44	0.1418	0.414	0.5849	0.378	0.6333
rs1474839	A	0.179	0.196	0.4458	0.207	0.225	0.182	0.9272
rs1884578	G	0.269	0.244	0.3423	0.251	0.5035	0.246	0.5634
rs1973943	T	0.448	0.41	0.1983	0.367	0.007	0.393	0.2096
rs4846908	A	0.443	0.473	0.2889	0.422	0.4739	0.445	0.955
rs1998064	C	0.219	0.23	0.6426	0.193	0.285	0.209	0.7935
rs2281719	G	0.394	0.428	0.2264	0.436	0.1381	0.426	0.4501
rs4846919	A	0.151	0.12	0.1309	0.158	0.7197	0.144	0.8194
rs1264084/rs606587	G	0.337	0.328	0.7492	0.334	0.9332	0.349	0.7544
rs4846928	T	0.142	0.128	0.5059	0.157	0.4457	0.134	0.7954
rs7512360	A	0.441	0.419	0.4517	0.412	0.3102	0.432	0.8299
rs3811486	C	0.116	0.106	0.5669	0.126	0.6233	0.099	0.5216
rs901675	T	0.297	0.274	0.389	0.307	0.7034	0.351	0.1629
rs1471915	T	0.114	0.12	0.7413	0.131	0.3403	0.142	0.3005
rs2273967	A	0.253	0.247	0.8128	0.247	0.8125	0.283	0.4453
rs16851339	T	0.021	0.02	0.9044	0.011	0.2414	0.015	0.6559
rs1043908	C	0.121	0.138	0.3637	0.114	0.7081	0.142	0.4506
rs1043909	A	0.188	0.215	0.2331	0.237	0.0358	0.149	0.2333
rs7022	C	0.325	0.324	0.979	0.289	0.1835	0.362	0.3672

Table 5-9 GALNT2 UC sub phenotypic analysis

Haplotype	Control freq	E3 freq	p-value	E2 freq	p-value	E1 freq	p-value
Block 1							
AT	0.785	0.759	0.2591	0.8	0.5311	0.784	0.9651
AC	0.107	0.126	0.2968	0.116	0.6165	0.115	0.7773
TT	0.107	0.115	0.6536	0.084	0.1757	0.101	0.8213
Block 2							
G	0.787	0.791	0.8716	0.796	0.7002	0.791	0.9276
T	0.213	0.209	0.8716	0.204	0.7002	0.209	0.9276
Block 3							
AG	0.602	0.544	0.0386	0.617	0.5788	0.645	0.3007
GG	0.21	0.261	0.0299	0.182	0.2339	0.172	0.2862
AA	0.189	0.196	0.7605	0.201	0.5951	0.182	0.8496
Block 4							
C	0.838	0.798	0.0581	0.831	0.727	0.831	0.8241
G	0.162	0.202	0.0581	0.169	0.727	0.169	0.8241
Block 5							
TGG	0.312	0.275	0.1623	0.303	0.7252	0.291	0.5962
AGG	0.29	0.289	0.9912	0.283	0.7995	0.33	0.294
AAG	0.22	0.241	0.3685	0.21	0.6968	0.196	0.5114
AAA	0.178	0.194	0.4482	0.204	0.2374	0.182	0.8952
Block 6							
AC	0.544	0.577	0.2406	0.608	0.0245	0.583	0.3488
GT	0.261	0.246	0.5581	0.25	0.653	0.243	0.6327
AT	0.191	0.175	0.4774	0.14	0.0231	0.172	0.5635
Block 7							
GA	0.557	0.526	0.2801	0.579	0.4478	0.557	0.9971
AA	0.223	0.243	0.4197	0.228	0.8475	0.234	0.7758
AC	0.22	0.231	0.6315	0.193	0.2662	0.209	0.7707
Block 8							
AG	0.604	0.572	0.2463	0.564	0.1589	0.574	0.4712
GG	0.244	0.306	0.0137	0.276	0.2034	0.279	0.348
GA	0.15	0.122	0.1675	0.159	0.654	0.147	0.9161
Block 9							
A	0.663	0.672	0.7499	0.666	0.9337	0.651	0.7559
G	0.337	0.328	0.7499	0.334	0.9337	0.349	0.7559
Block 10							
CGTT	0.294	0.275	0.4645	0.306	0.6465	0.351	0.1471
CGTC	0.268	0.305	0.1393	0.282	0.585	0.216	0.1734
CATC	0.181	0.182	0.9694	0.136	0.0387	0.205	0.4825
TATC	0.139	0.129	0.6214	0.154	0.4356	0.128	0.7245
CACC	0.116	0.108	0.6575	0.122	0.7869	0.1	0.5418
Block 11							
CGA	0.642	0.639	0.9143	0.633	0.7348	0.595	0.2418
CAA	0.223	0.22	0.9023	0.224	0.9524	0.246	0.5046
TGA	0.105	0.114	0.6022	0.121	0.3371	0.125	0.4202
CAT	0.021	0.021	0.958	0.011	0.2142	0.017	0.7413
Block 12							
TGG	0.486	0.46	0.3653	0.474	0.6849	0.486	0.9968
TGC	0.205	0.187	0.4437	0.177	0.2138	0.223	0.5948
TAG	0.188	0.214	0.23	0.234	0.0411	0.148	0.2365
CGC	0.118	0.137	0.3226	0.111	0.6924	0.135	0.5564

Table 5-10 GALNT2 UC sub phenotypic haplotype analysis

5.5 WTCCC analysis

The publicly available WTCCC SNP data across the GALNT2 gene were downloaded from the WTCCC website by Dr Gail Davies. As the Edinburgh cohort formed part of the WTCCC study these samples were excluded from this analysis in order to prevent some samples being in both the WTCCC analysis and the separate Edinburgh analysis of the Illumina® data. Information downloaded included the p-values for each SNP but not the minor allele frequencies. The p-values were plotted against the SNP location. There were 92 WTCCC SNPs, thus to correct for multiple testing across the gene a p-value <5×10^{-4} was required. Gene-wide rather than genome-wide significance was considered appropriate for a gene which already had a hypothesis for CD susceptibility. The results are shown in Figure 5-3. Only one marker reached gene-wide significance: rs12751815 (p-value 1.85×10^{-4}). Two further SNPs, rs1358769 and rs7513659, approached significance (p-values 0.001 each). There was a cluster of SNPs between exons 1 and 2 that had smaller p-values than in other areas, giving the suggestion that it may be an area of significance.

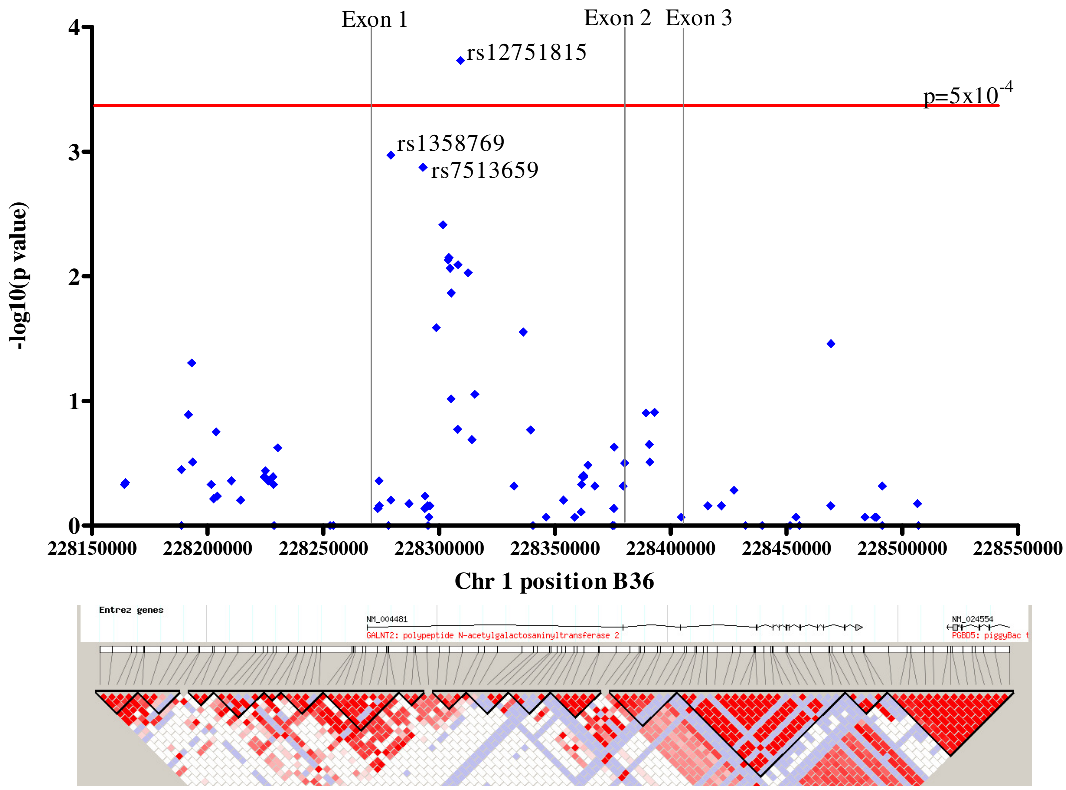


Figure 5-3 Non Scots WTCCC GALNT2 data and Haploview map of the region

5.6 Replication of findings

Although the WTCCC data analysis was a reassuring 'surrogate' replication, it was necessary to try to get replication of the association between CD and rs2281719 in a separate cohort of Scottish patients. At the time of requesting this genotyping, the Dundee cohort was incomplete and no controls were available, so the purpose of genotyping was to attempt replication of the MAF in cases. In the Edinburgh cohort the MAF of rs2281719 was 0.45 in IBD and 0.47 in CD, with a control MAF of 0.39. However, in the limited Dundee cohort of 155 CD, 119 UC, the MAF was 0.38, with no difference between UC and CD.

These results were puzzling and hard to explain. The WTCCC dataset was re-examined, and it appeared that rs2281719 had been genotyped in this population, but without showing any significant association when comparing WTCCC CD cases against controls (p-value 0.327, MAF not available).

SNP calls on the Illumina® platform are automatically generated by the Beadstudio clustering algorithm. The Illumina® cluster plot was examined (Figure 5-4) for suboptimal clustering of rs2281719, and it was felt that reclustering would be appropriate as there was poor separation between the wild type homozygotes (red) and the heterozygotes (purple) samples. The reclustering was done manually by Angie Fawkes in the WTCRF, Edinburgh. However, on reanalysis of the SNP, the MAF for cases remained 0.46 for IBD, and 0.39 for controls.

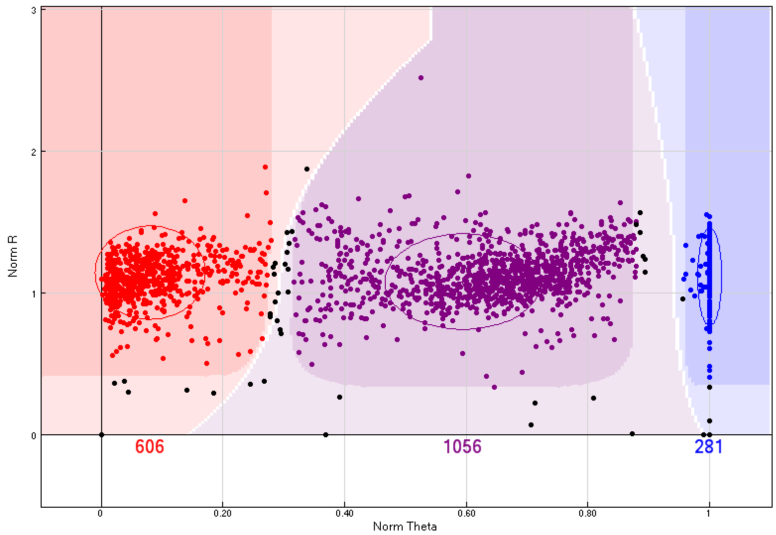


Figure 5-4 rs2281719 before reclustering

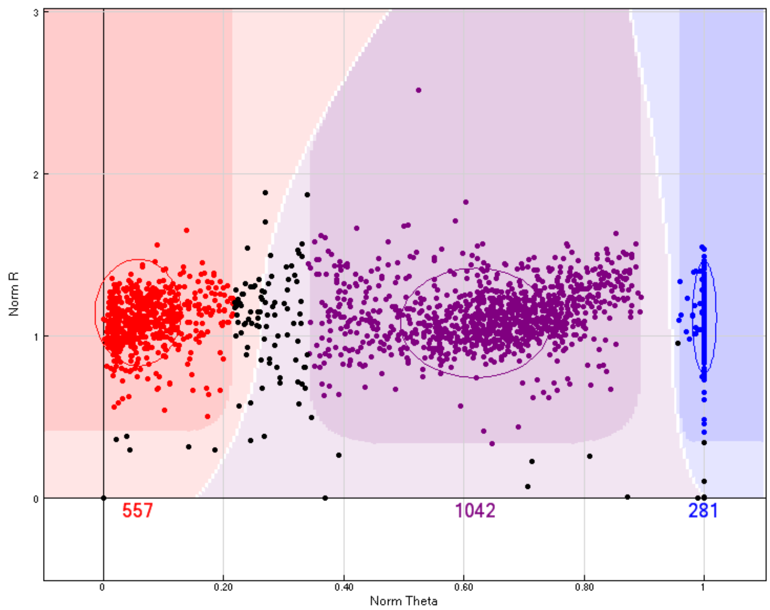


Figure 5-5 rs2281719 after reclustering

The next step was to sequence across the SNP in specific patients to check the accuracy of the Illumina® calling. The primers used were 5'ATACATGAAGTCTCTGGGCC3' (forward) and 5'ATATGGTAGGTGATCAGATACAGG 3'(reverse) (primers 763/764), which were optimized for an annealing temperature of 62°C with an initial gradient PCR. A total of 63 DNAs were selected: 36 GA, 19 AA and 8 GG samples for sequencing across the rs2281719 SNP. The sequencing results matched the genotyping results in all of the AA cases. Of the 8 GG genotyping, 7 were GG when sequenced, and one was AA. Of the 36 supposed heterozygotes, 29 correlated with the sequencing, but 7 did not, and appeared to be AA on the basis of the sequencing. This low correlation of Illumina® genotyping with the sequencing data, particularly for the heterozygotes, was of major concern.

The next step was to perform Taqman® genotyping for rs2281719, in order to clarify the situation for as many of the dataset as possible. As the DNA plates that had been used for both Illumina® and Taqman® genotyping had run low, the DNA had been replated meaning that not all of the samples genotyped on the Illumina® platform were genotyped on Taqman®. Of note, on Taqman® the MAF in both CD cases and controls was 0.37. In total there were 995 samples with both Illumina® and Taqman® genotyping; the results of this analysis are shown in Table 5-11.

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="4">Taqman® seq</td></tr>
<tr><td>Not called</td><td>GG</td><td>AG</td><td>AA</td></tr>
<tr><td rowspan="4">Illumina® seq</td><td>Not called</td><td>0</td><td>1</td><td>4</td><td>3</td></tr>
<tr><td>GG</td><td>10</td><td>15</td><td>4</td><td>0</td></tr>
<tr><td>AG</td><td>16</td><td>3</td><td>522</td><td>95</td></tr>
<tr><td>AA</td><td>10</td><td>2</td><td>8</td><td>302</td></tr>
</table>

Table 5-11 rs2281719 Illumina® and Taqman® sequencing compared

In total there were 112 discrepancies between the datasets, representing 11.3% of the samples, including 2 samples that had been called on Illumina® as wild type homozygous which on Taqman® were called minor allele homozygous. One or other of the genotyping platforms was seriously flawed for this genotyping.

In order to clarify which one of the 2 genotyping platforms was incorrect for the genotyping of this SNP, as many as possible of the discrepant samples were sequenced using primers 763/764, as detailed earlier in this section. In total it was possible to sequence 83 of these samples. In all of these cases the sequencing correlated with Taqman® rather than Illumina®, including one of the samples that had been called on Illumina® as AA but was called as GG on Taqman®. When these samples were looked at on the Illumina® clustering, only 54 of the samples were in areas that could be no-calls if the clustering on Illumina® was made more stringent, as shown in Figure 5-6.

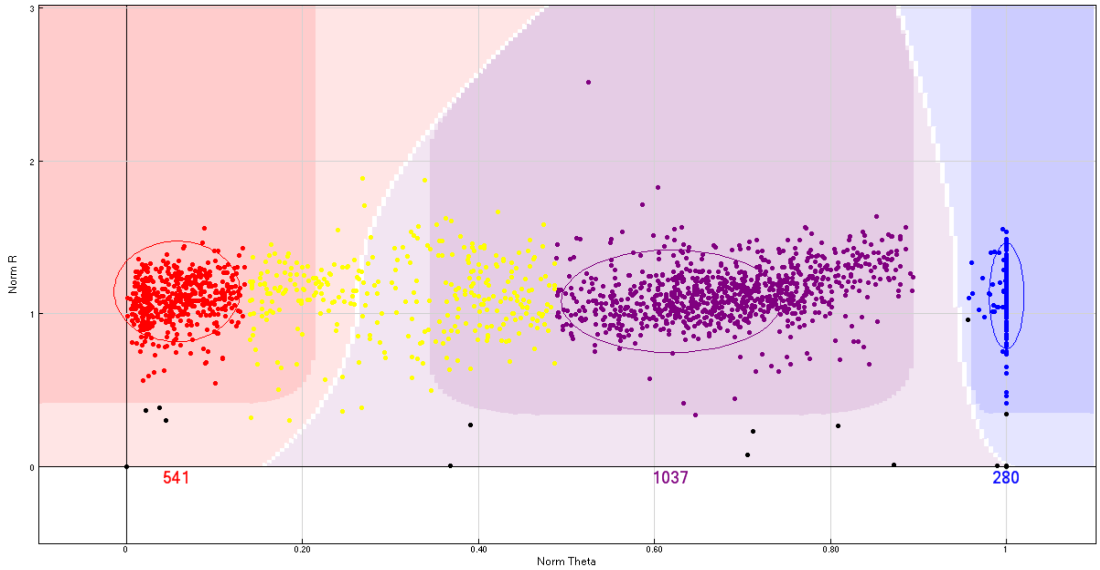


Figure 5-6 rs2281719 Illumina® cluster plot with more stringent clustering

The rest of the discrepant samples (29) fell within the centre of the clusters. So not only were more than 10% of the samples not called on the more severe clustering (thus not attaining pre-determined quality control levels) but in addition some of the samples appeared to have been incorrectly called. It was not clear whether this was a problem with the SNP genotyping or a problem with the plates on which the genotyping had been done. Advice was taken from Dr Mark Gibbs, an Illumina® field application employee. After looking at the data for rs2281719, his response was:

> 'The SNP itself is on the 1M chip, and is in dbSNP with quite a lot of data. Allele frequencies for the SNP for Caucasians are 0.35, 0.483 and 0.187. Relative to these your data has an excess of AA and AG calls, and a lack of GG calls, and this is a significant difference (according to my back of envelope chi square test). All in all it looks like this Goldengate assay is not resolving the genotypes, which can happen if there are additional polymorphisms, other similar sequences in the genome etc that either affect the binding of primers, amplification of alleles and so on creating null alleles or unusual, these kinds of things could also affect the Taqman® assay of course. In this sense, Infinium is

> more robust than Goldengate I think because it hinges on extension from a single long primer. I don't know if this is significant, but the ASO sequence has two perfect matches in the genome, the correct on Chr 1, and another on chr12, even if the chr 12 match won't make a PCRable product because the LSO is a perfect match only to Chr1, perhaps there is some variation in signal introduced.'

In view of this knowledge, data from rs2281719 had to be discounted. As a large number of the controls had not been sequenced on Taqman®, it was not possible to incorporate the Taqman® data with the remaining *GALNT2* genotyping.

This concern about rs2281719 led us to re-call and check all the data generated on the Illumina® platform. A large number of SNPs (1536) had been genotyped covering not only the yeast-2 hybrid candidates but also fine mapping the IBD2 region (25Mb) on chromosome 12. Angie Fawkes at the Genetics Core of the WTCRF examined the other SNPs that had been genotyped on the Illumina® platform at the same time and on a number had similarly questionable clustering. One of the problems appeared to be differences in sample concentration between plates. Angie Fawkes went through each SNP in the dataset and reclustered each one manually, calling every plate separately. This generated a new set of *GALNT2* data, which had to be completely reanalysed.

A number of SNPs from the other genes that had been genotyped on the Illumina® platform were also genotyped on the Taqman® platform, and this genotyping correlated well with the reclustered data, giving us confidence in the Illumina® platform for the rest of the sequencing.

5.7 GALNT2 reanalysis

5.7.1 Quality control

Initial quality control removed only one SNP in addition to rs2281719: rs16851339, due to poor genotyping (following reclustering of all the SNPs). Interestingly, rs868287, despite having been excluded on the initial analysis (see section 5.4.1), was not removed on the reanalysis. This was as a result of more DNA samples being excluded and the poor genotyping having been with those excluded samples. The numbers of samples analysed were: 351 CD, 434 UC and 841 controls as 23 CD, 36 UC and 37 control samples were removed at the quality control stage.

5.7.2 GALNT2 reanalysis - single SNP analysis

When corrected for multiple testing (p-value<0.002), the SNP rs7536663 (located in intron 1) was the only single SNP demonstrating association (OR 1.38, 95% CI 1.14-1.67, p-value $8x10^{-4}$), as shown in Table 5-12.

	Minor allele	Control MAF	IBD MAF	p-value	CD MAF	p-value	UC MAF	p-value
rs12135308	T	0.101	0.109	0.441	0.121	0.1633	0.1	0.9465
rs7515331	C	0.106	0.117	0.3433	0.123	0.2299	0.111	0.7003
rs1474925	T	0.213	0.205	0.5408	0.204	0.6031	0.205	0.6273
rs910820	G	0.206	0.221	0.3322	0.233	0.1577	0.21	0.8319
rs11122300	A	0.186	0.183	0.8225	0.176	0.6172	0.188	0.8873
rs12029443	G	0.162	0.163	0.9078	0.148	0.4321	0.176	0.3629
rs7536663	T	0.281	0.314	0.0407	0.351	8.00E-04	0.284	0.8629
rs2057233	A	0.401	0.385	0.3714	0.351	0.0273	0.413	0.5517
rs868287	C	0.283	0.296	0.395	0.293	0.6346	0.299	0.3843
rs1474839	A	0.178	0.176	0.8682	0.154	0.1707	0.194	0.3319
rs1884578	G	0.26	0.276	0.3039	0.311	0.0135	0.248	0.5198
rs1973943	T	0.462	0.458	0.8558	0.501	0.0807	0.422	0.0634
rs4846908	A	0.441	0.447	0.7419	0.461	0.3595	0.435	0.7592
rs1998064	C	0.222	0.234	0.3819	0.267	0.0177	0.208	0.4201
rs4846919	A	0.153	0.149	0.7562	0.156	0.8747	0.144	0.5311
rs1264084/ rs606587	G	0.335	0.335	0.9825	0.337	0.9168	0.334	0.9598
rs4846928	T	0.145	0.15	0.6677	0.157	0.4487	0.144	0.979
rs7512360	A	0.447	0.427	0.257	0.436	0.6182	0.42	0.1963
rs3811486	C	0.118	0.114	0.7436	0.117	0.9376	0.112	0.6598
rs901675	T	0.291	0.288	0.8597	0.262	0.1457	0.31	0.3239
rs1471915	T	0.115	0.12	0.6597	0.104	0.4382	0.133	0.1882
rs2273967	A	0.257	0.258	0.9785	0.25	0.7283	0.264	0.7219
rs1043908	C	0.121	0.122	0.9505	0.102	0.2007	0.138	0.2341
rs1043909	A	0.186	0.199	0.3418	0.19	0.8164	0.207	0.2085
rs7022	C	0.328	0.322	0.7374	0.32	0.7225	0.324	0.8494

Table 5-12 GALNT2 reanalysis - single SNP analysis

5.7.3 GALNT2 reanalysis - Haplotype analysis

A haplotype analysis was completed, defining the haplotype block according to solid spine of LD on the control samples only (Figure 5-7). The results are shown in Table

5-13. Haplotypes in block 5 were associated with CD susceptibility: the TGGG haplotype in intron 1 showed association with CD (p = 0.0017). The AGGG haplotype had a higher frequency in controls rather than cases, but the frequencies in the cohort were very low, meaning that it could represent an artifact.

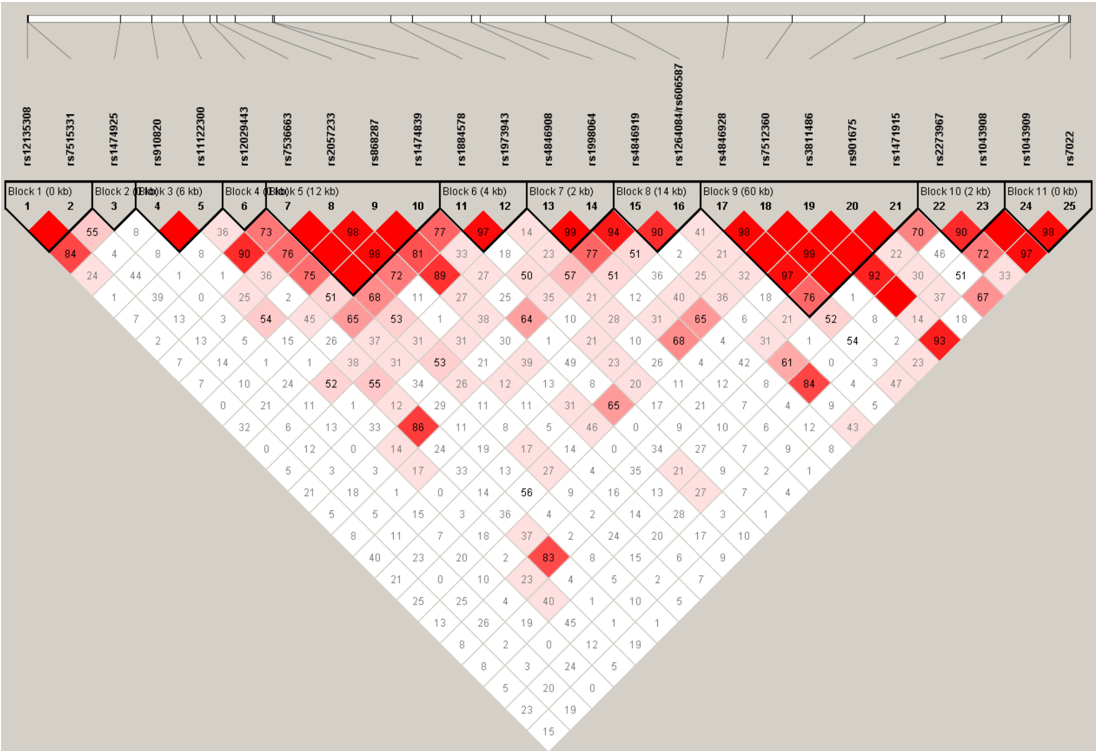


Figure 5-7 GALNT2 reanalysis haplotypes, controls only

Haplotype	Controls	IBD	p-value	CD	p-value	UC	p-value
Block 1							
AT	0.793	0.774	0.2031	0.757	0.0537	0.789	0.8186
AC	0.106	0.116	0.3447	0.123	0.2314	0.111	0.7018
TT	0.101	0.109	0.4593	0.12	0.1738	0.1	0.9325
Block 2							
C	0.787	0.795	0.5433	0.796	0.6055	0.795	0.6293
A	0.213	0.205	0.5433	0.204	0.6055	0.205	0.6293
Block 3							
AG	0.608	0.597	0.5243	0.591	0.4239	0.602	0.7705
GG	0.207	0.22	0.3346	0.233	0.1539	0.21	0.8434
AA	0.185	0.182	0.8291	0.176	0.6146	0.188	0.8726
Block 4							
C	0.838	0.837	0.9084	0.852	0.4348	0.824	0.3672
G	0.162	0.163	0.9084	0.148	0.4348	0.176	0.3672
Block 5							
TGGG	0.288	0.315	0.0873	0.352	0.0017	0.284	0.8901
AGCG	0.281	0.295	0.3852	0.287	0.77	0.302	0.3057
AAGG	0.223	0.208	0.2906	0.196	0.1406	0.218	0.7346
AAGA	0.176	0.176	0.9829	0.153	0.1848	0.195	0.2671
AGGG	0.029	0.003	1.39E-08	0.007	7.00E-04	0.001	1.07E-06
Block 6							
TC	0.536	0.539	0.8627	0.498	0.0982	0.572	0.08
CT	0.261	0.276	0.3157	0.312	0.0115	0.247	0.4774
TT	0.201	0.184	0.2414	0.19	0.5318	0.18	0.2022
Block 7							
GA	0.559	0.554	0.7899	0.54	0.3975	0.566	0.7354
AC	0.221	0.235	0.3389	0.268	0.0141	0.208	0.4499
AA	0.22	0.211	0.5137	0.192	0.1288	0.226	0.7301
Block 8							
GA	0.656	0.658	0.8877	0.657	0.9678	0.659	0.8629
GG	0.191	0.193	0.8448	0.188	0.916	0.198	0.6821
AG	0.144	0.143	0.9191	0.15	0.7167	0.137	0.6231
Block 9							
CGACC	0.267	0.285	0.2664	0.301	0.0894	0.271	0.8226
CAACC	0.181	0.163	0.1882	0.163	0.2986	0.164	0.2817
CGATC	0.176	0.168	0.5308	0.158	0.2725	0.176	0.9968
TAACC	0.143	0.149	0.6267	0.157	0.3833	0.142	0.972
CGATT	0.115	0.12	0.6125	0.105	0.4673	0.134	0.1692
CAGCC	0.119	0.115	0.76	0.117	0.92	0.113	0.6977
Block 10							
CT	0.735	0.736	0.9471	0.744	0.6685	0.73	0.7768
TT	0.144	0.142	0.9274	0.155	0.5074	0.133	0.4588
TC	0.113	0.115	0.8663	0.096	0.2424	0.131	0.2089
Block 11							
GG	0.487	0.48	0.7144	0.493	0.7946	0.469	0.4044
GC	0.327	0.32	0.6879	0.316	0.6263	0.324	0.8487
AG	0.185	0.197	0.3722	0.186	0.8997	0.207	0.1994

Table 5-13 GALNT2 reanalysis - haplotype analysis

5.7.4 Sub phenotypic reanalysis – CD

Results of an analysis of genotype with CD sub phenotype are shown in Table 5-14. No values reached statistical significance, although rs7536663 had a borderline significant association with L1 disease (p-value of 0.0024). On haplotypic analysis no values reached statistical significance (data not shown). The TGGG haplotype in block 5 had a non significant p-value of 0.004 in L1 disease.

SNP	Minor allele	Control MAF	L1 MAF	p-value	L2 MAF	p-value	L3 MAF	p-value	B3 MAF	p-value	B2 MAF	p-value	B1 MAF	p-value
rs12135308	T	0.101	0.12	0.3968	0.116	0.4887	0.093	0.7523	0.114	0.6618	0.123	0.4807	0.122	0.1973
rs7515331	C	0.106	0.119	0.5713	0.138	0.1475	0.099	0.7768	0.184	0.0104	0.085	0.4877	0.117	0.4974
rs1474925	T	0.214	0.21	0.9112	0.196	0.5286	0.152	0.0881	0.224	0.7891	0.208	0.8828	0.198	0.4712
rs910820	G	0.207	0.274	0.0255	0.24	0.2406	0.214	0.8323	0.295	0.0276	0.222	0.7005	0.221	0.5039
rs11122300	A	0.185	0.154	0.2774	0.183	0.9616	0.193	0.8098	0.175	0.8066	0.186	0.9668	0.174	0.5886
rs12029443	G	0.161	0.15	0.6694	0.158	0.9198	0.143	0.5757	0.198	0.2921	0.142	0.5975	0.137	0.2079
rs7536663	T	0.281	0.382	0.0024	0.344	0.0432	0.343	0.1269	0.377	0.0288	0.365	0.0659	0.342	0.0112
rs2057233	A	0.4	0.311	0.0128	0.36	0.2302	0.407	0.8663	0.362	0.4212	0.324	0.1185	0.355	0.0773
rs868287	C	0.283	0.294	0.727	0.298	0.639	0.254	0.4538	0.241	0.3352	0.306	0.6139	0.302	0.4147
rs1474839	A	0.177	0.145	0.2361	0.154	0.3741	0.196	0.592	0.202	0.5127	0.144	0.3875	0.145	0.0961
rs1884578	G	0.26	0.329	0.0362	0.298	0.2152	0.312	0.1896	0.351	0.0347	0.324	0.1599	0.299	0.0995
rs1973943	T	0.462	0.486	0.1526	0.483	0.5374	0.521	0.1765	0.588	0.0093	0.481	0.6954	0.485	0.3711
rs4846908	A	0.441	0.457	0.6551	0.459	0.5947	0.486	0.3001	0.412	0.5515	0.435	0.9078	0.479	0.1392
rs1998064	C	0.222	0.262	0.1896	0.275	0.0667	0.286	0.0829	0.246	0.5551	0.288	0.1148	0.268	0.0359
rs4846919	A	0.153	0.173	0.4531	0.127	0.2863	0.196	0.1867	0.211	0.1032	0.142	0.7466	0.146	0.694
rs1264084/rs606587	G	0.335	0.36	0.4684	0.32	0.6367	0.338	0.9373	0.33	0.921	0.34	0.9208	0.338	0.894
rs4846928	T	0.145	0.17	0.3376	0.145	0.989	0.134	0.7164	0.132	0.694	0.17	0.4833	0.16	0.4027
rs7512360	A	0.447	0.435	0.733	0.446	0.9749	0.4	0.2837	0.431	0.7396	0.417	0.5403	0.441	0.8248
rs3811486	C	0.118	0.118	0.988	0.103	0.4972	0.141	0.427	0.121	0.9381	0.104	0.6528	0.119	0.954
rs901675	T	0.291	0.257	0.2945	0.25	0.1808	0.296	0.9137	0.293	0.9701	0.269	0.6106	0.253	0.0997
rs1471915	T	0.115	0.113	0.9286	0.103	0.5831	0.121	0.8274	0.149	0.2777	0.12	0.8729	0.91	0.118
rs2273967	A	0.257	0.233	0.4601	0.269	0.6862	0.269	0.7644	0.282	0.5629	0.212	0.3028	0.251	0.7992
rs1043908	C	0.12	0.094	0.2676	0.095	0.2573	0.106	0.6256	0.118	0.9458	0.115	0.8796	0.095	0.127
rs1043909	A	0.186	0.205	0.5069	0.158	0.3064	0.187	0.9821	0.182	0.9175	0.212	0.5136	0.187	0.9543
rs7022	C	0.328	0.308	0.5702	0.321	0.8255	0.291	0.383	0.286	0.3581	0.346	0.698	0.323	0.8359

Table 5-14 GALNT2 CD sub phenotypic reanalysis

5.7.5 Sub phenotypic reanalysis – UC

Results of a sub phenotypic analysis for *GALNT2* in UC are not shown, but did not demonstrate any statistically significant association with UC susceptibility either with single markers or haplotypes.

5.7.6 Replication

The positive result for rs7536663 merited an attempt at replication and was genotyped on the Taqman® platform in the Dundee cohort of 300 CD, 211 UC patients and 682 controls. The results are given in Table 5-15. Although there was a trend towards a difference in allele frequencies in CD compared with controls, it was far from being statistically significant, unlike the Edinburgh MAF of 0.351 and 0.281 in CD and controls respectively.

		MAF	p-value vs controls
Controls	rs7536663T	0.309	
IBD	rs7536663T	0.312	0.874
CD	rs7536663T	0.326	0.469
UC	rs7536663T	0.293	0.5319

Table 5-15 Dundee genotyping of rs7536663

5.8 GALNT2 exonic sequencing

At the point of finding the positive results for rs2281719 in CD patients and the WTCCC association in the same area of the gene (the intronic area between exons 1 and 2) mentioned in sections 5.4 and 5.5, all 16 exons of the GALNT2 gene were sequenced in patients with specific haplotypes in an attempt to find potential causative mutations. As the GG haplotype in block 8 in CD patients had demonstrated a lower p-value compared with rs2281719, DNA was selected according to this haplotype rather than solely on the basis of the rs2281719 genotype.

Numbers of DNAs sequenced are shown in Table 5-16. By sequencing 25 of the GG/GG risk haplotype there was sufficient power to detect mutations occurring at a rate of at least 4%. There were no patients with the AA/AA haplotype.

Haplotype	rs2281719	rs4846919	Control DNA	CD DNA	Total
1	AA	GG	73	53	126
2	GG	AA	4	5	9
3	GG	GG	11	14	25

Table 5-16 GALNT2 Exonic sequencing: DNA selection

5.8.1 GALNT2 Exon 1 PCR

GALNT2 exon 1 proved problematic to amplify. An initial PCR with primers 804/805 failed, and despite doubling the amount of Taq polymerase there was no band on the post-gradient PCR gel. A further PCR with increased DNA, magnesium and Taq worked, and so the DNAs were taken to the HGU MRC Technical services department for sequencing. However when the sequencing was analysed it was clear that, although the sequencing was clean, the post-PCR product was not *GALNT2* exon 1. On review of the gradient PCR that had appeared to work it was clear that the product size was actually too small (Figure 5-8), with a product size of 300bp when the expected size was about 900bp.

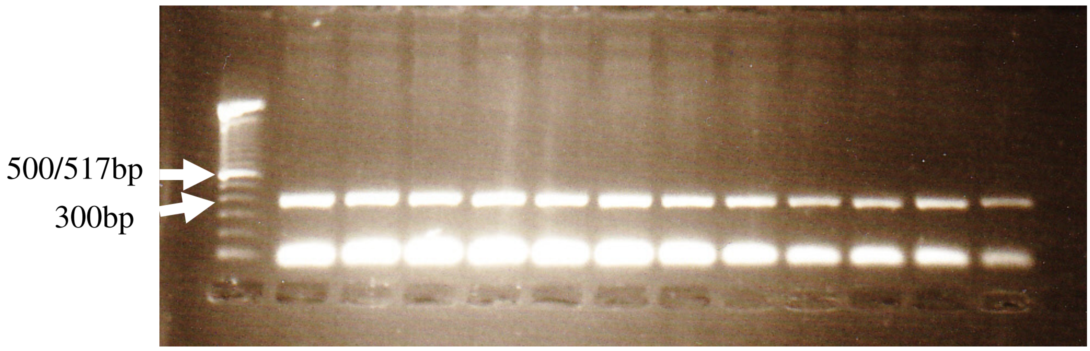


Figure 5-8 GALNT2 Exon 1 Gradient PCR E57D10

A second set of primers were designed (853/854) but that PCR did not work: the post-PCR gel was blank despite several attempts with varying quantities of Taq, magnesium and DNA, and a 1 minute extension to each PCR cycle. Combinations of primers (804/854 and 853/805) were also attempted. There was a band of the correct size on the post PCR gel with primer combinations 804/854 (Figure 5-9), but there were non specific DNA fragments present, which despite optimizing with varying Taq and magnesium, could not be improved.

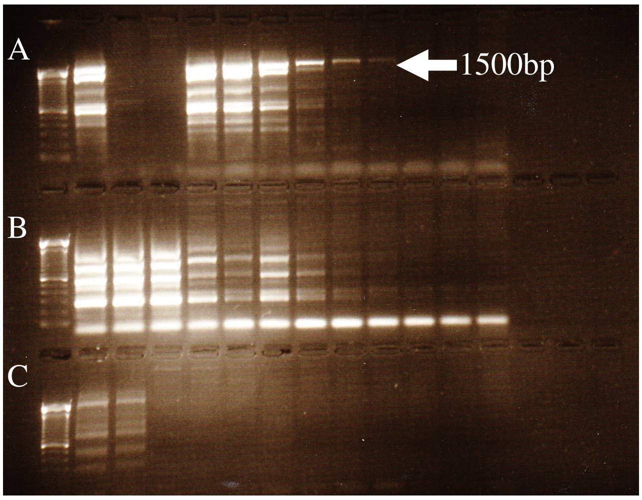


Figure 5-9 Gradient PCR E60D10 with 1 minute extension: A 804/854 with correct product size (1500bp) and non specific binding; B 853/854 expected product size 900bp - did not work; C 853/805 expected product 700bp – did not work

Before proceeding to full sequencing with this primer set and conditions, a small number of DNAs were sequenced with a nested PCR. The optimized conditions were used with primers 804/854 for the PCR, and primers 853/805 were used for the sequencing reaction (Figure 5-10). Unfortunately the resulting sequencing was extremely poor, with unreadable sequence in the centre (Figure 5-11).

Figure 5-10 Exon 1 primers; exon highlighted in red

Figure 5-11 Exon 1 Trial sequencing

Due to the high GC content of the region with the resultant possibility of tight DNA folding preventing the primers from accessing the relevant DNA, a commercially available Taq polymerase, specific for high GC content PCRs was used (BioXAct Short, Bioline) at a working concentration of 0.08 units/µl, along with Polymate Additive (a melting agent to improve reaction specificity, Bioline) at the manufacturer's recommended concentration. Bioline's 10x reaction buffer and $MgCl_2$ were also used (Bioline). The PCR run was according to Bioline's protocols, and was different to the conditions used in standard PCRs, as shown in Table 5-17.

95°C for 5 minutes (initial denaturation)	
Target temperature for 1 minute (annealing)	
70°C for 2 minutes (extension) 95°C for 30 sec (denaturation) Target temp for 30 sec (annealing)	}repeated }30 }times
70°C for 10 minutes (final elongation)	

Table 5-17 Bioline PCR conditions

Unfortunately the post PCR gels remained blank, despite trying different primer combinations (853/854, 804/854, 853/805) and varying the amounts of Taq, magnesium and DNA.

Finally Betaine (trimethylglycine, Sigma) at a working concentration of 0.1M was used, and the resulting PCRs all worked, producing a product of the expected size (Figure 5-12). Betaine, by mechanisms that are not understood, helps to prevent the formation of secondary structures in DNA. The best primer combination (853/805) was repeated with several DNAs and the normal in-house Taq and conditions, but with 0.1M Betaine. This PCR also worked well, and sequencing of 8 DNAs produced good quality sequencing (Figure 5-13).

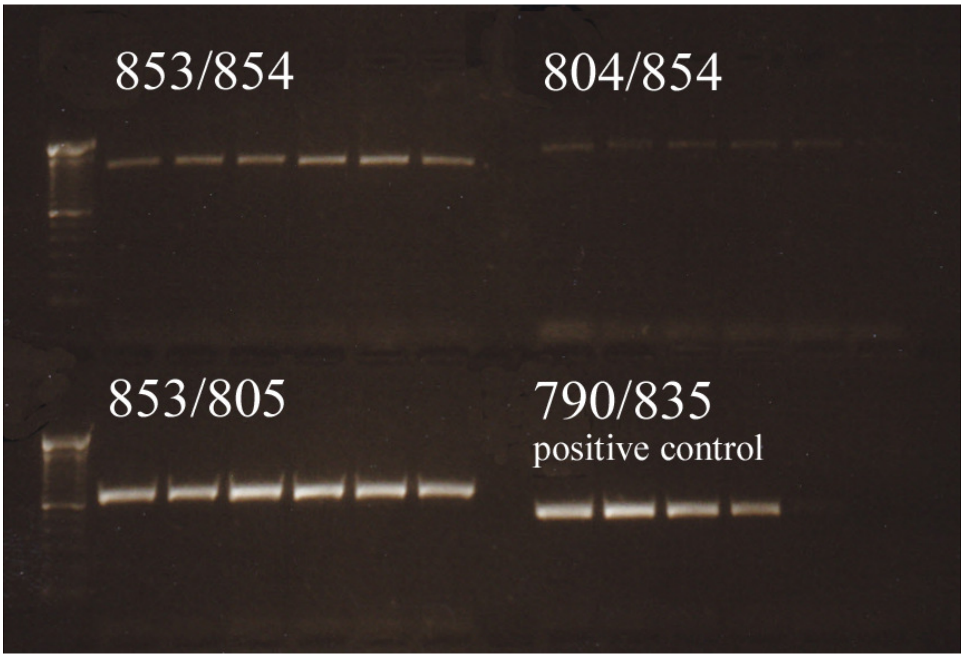


Figure 5-12 Gradient PCRs run on E60D10M2 programme

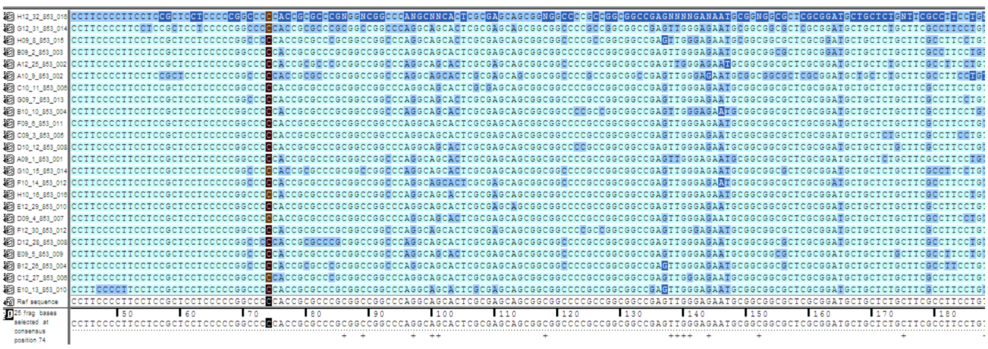


Figure 5-13 Check sequencing 853/805; black line is at beginning of Exon 1

5.8.2 GALNT2 sequencing results

Once the PCRs were optimized for exons 1-16, PCRs were completed on the DNAs as shown in Table 5-16. The results were analysed according to genotype. Each exon was examined, including, where possible, 50bp up and downstream of the exon itself. Not all the samples demonstrated good sequencing, as shown in Table 5-18.

	50bp upstream (%)	Exon (%)	50bp downstream (%)
Exon 1	73.8	39.3	39.3
Exon 2	64.3	64.3	64.3
Exon 3	65.5	72.6	71.4
Exon 4	4.8	37.5	35.7
Exon 5	36.3	36.3	36.3
Exon 6	37.5	37.5	20.2
Exon 7	57.7	56.5	56.5
Exon 8	59.5	58.9	58.9
Exon 9	39.9	39.9	39.9
Exon 10	64.9	64.9	64.9
Exon 11	50.6	50.6	50.6
Exon 12	79.7	82.1	82.1
Exon 13	81.5	81.5	81.5
Exon 14	63.1	63.1	63.1
Exon 15	67.2	66.6	66.6
Exon 16	47.6	51.8	51.8

Table 5-18 Percentage successful sequencing for each GALNT2 exon and immediately adjacent intronic areas

There were problems with the quality of the sequencing of particular exons, especially exons 4-6. However, there were still enough good sets of sequencing to allow a meaningful analysis of each area. A summary of the analysis is shown in Table 5-19 which only lists the 15 samples in which mutations were found. Within the areas in and around exons 1 and 2, ie around the areas of haplotype block 8, there was one intronic SNP (rs678050) about 50bp upstream of exon 2 which was homozygous for the rare allele in 2 DNAs: 040CD (case) and 204BT(control). There were 2 synonymous coding SNPs in exon 9, rs3748006 and rs1923950. One of the sequenced samples (340CD) was heterozygous for rs3748066 and homozygous for the wild type genotype with rs1923950. Another two samples (040CD and 306CD) were heterozygous for rs1923950 and wild type homozygous for rs3748006. The only non synonymous SNP found was in Exon 16 (rs2273970) but the heterozygote was only found in one DNA (340CD), as all the other samples that were successfully sequenced were wild type homozygous. In conclusion, the sequencing did not reveal any new non synonymous SNPs likely to be correlated with the genotype of haplotype block 8.

			Exon 2	Exon 4	Exon 5	Exon 9		Exon 10		Exon 12	Exon 15	Exon 16		
SNP wt/minor allele SNP Type	rs2281719 A/G Intronic	rs4846919 A/G Intronic	rs678050 T/C Intronic	rs3811488 G/A Intronic	rs3811486 A/G Intronic	rs3748006 T/C Syn Coding	rs1923950 G/A Syn Coding	rs2273965 C/T Intronic	rs2273966 T/A Intronic	rs3811485 G/A Intronic	rs12091838 C/T Intronic	rs6698963 A/G Intronic	rs2273970 G/A Non Syn	rs2273969 C/T 3' UTR
040CD	AA	GG	CC	-	-	wt	GA	wt	TA	wt	CT	AG	wt	wt
215CD	AA	GG	-	GA	AG	-	-	-	-	GA	-	-	-	-
264CD	AA	GG	-	AA	wt	-	-	wt	wt	wt	-	-	-	-
280CD	AA	GG	wt	-	-	-	-	-	-	-	wt	AG	wt	wt
296CD	AA	GG	-	-	-	-	-	-	-	wt	-	-	-	-
306CD	AA	GG	wt	-	-	wt	GA	wt	TA	wt	CT	AG	wt	wt
329CD	AA	GG	wt	-	-	-	-	-	-	-	wt	wt	wt	wt
333CD	AA	GG	wt	-	-	-	-	wt	TA	wt	wt	wt	wt	wt
340CD	AA	GG	wt	-	-	TC	wt	CT	wt	wt	wt	AG	GA	wt
204BT	GG	GG	CC	-	-	-	-	wt	AA	AA	wt	-	-	-
231BT	GG	GG	wt	-	-	-	-	-	-	wt	wt	-	-	-
007HC	GG	GG	-	GA	wt	-	-	wt	wt	wt	-	-	-	-
052HC	GG	GG	wt	wt	wt	wt	wt	wt	wt	wt	wt	AG	wt	CT
084HC	GG	GG	wt	-	-	-	-	wt	wt	wt	wt	-	-	-
121HC	GG	GG	wt	GA	wt	wt	wt	wt	TA	GA	wt	AG	wt	CT

Table 5-19 GALNT2 exon sequencing, only samples that had mutations are listed here, wt=wild type, -=not sequenced

5.8.3 Further Sequencing

The DNAs used in the sequencing across the *GALNT2* gene detailed in section 5.8.2 had been chosen on the basis of a haplotype that was subsequently found not to be valid. Therefore further sequencing was completed on *GALNT2* exons 1 and 2 on the basis of the TGGG haplotype detailed in section 5.7.3. Only these exons were sequenced because the TGGG haplotype lay between these exons. Twelve CD patients with the 'at risk' TGGG haplotype and twelve controls with at least one chromosome having the AGGG haplotype (the other chromosome having either the AAGG or the AGCG haplotype) were selected. These samples were amplified by PCR with the primers and conditions for exons 1 and 2 listed in section 5.3.2, and sequenced at the MRC Human Genetics Unit, Edinburgh.

5.8.3.1 Results

Of the 12 CD samples, 9 produced good sequencing across the whole of exon 1, two samples completely failed and one other had poor sequencing for the 50bp upstream of the exon, but good sequencing otherwise. Of the 12 control samples, 8 produced good sequencing, 3 completely failed and one sample had poor sequencing for the 50bp upstream of the exon, but good sequencing otherwise. No SNPs were found in this region in the samples that had been successfully sequenced.

All 24 samples produced good sequencing for exon 2, and no SNPs were found in this exon and the regions 50bp upstream and downstream of the exon.

5.9 Discussion

This chapter has demonstrated that rs7536663 in the *GALNT2* gene is associated with CD susceptibility in the Edinburgh cohort with an OR of 1.38. There was a trend to a difference in the MAF in CD and controls in the Dundee replication cohort but the replication cohort was underpowered to rule out an association: with an OR of 1.38 a minimum of 695 CD cases in the replication cohort would be needed in order to confidently rule out an association with CD, and this number rises to 1035 with an OR of 1.3, and 2110 with an OR of 1.2. Other CD cohorts from other groups could also be used for this purpose but then the population specific nature of the potential genetic association would be lost. Recent WTCCC immunochip genotyping has included the rs7536663 SNP; results are eagerly awaited. Of course, the question of what is a significant p-value is a constant concern in hypothesis-driven genetic research. There have been issues with replicated genetic loci from candidate gene studies not being uncovered in GWAS or meta-analyses; *DLG5* is a good example of this.[145] It is not clear whether this lack of association in GWAS is due to population pooling diluting out a population-specific gene or whether it was really a false positive.

GWAS are an important tool in the search for susceptibility genes in complex diseases. However, they have their limitations. Not only are they unable to uncover genes which have low OR for susceptibility, but they are also not able to distinguish population specific susceptibility genes due to the need for large populations and the necessary pooling of cohorts. The yeast two-hybrid screen was a plausible alternative to help negate the second problem of population specific susceptibility genes. As *NOD2* is a known susceptibility gene with a low rate of mutations in the Scottish cohort, it made a sensible bait protein for the screen, to try to find other genes in its pathways that may prove to be susceptibility genes as well.

The amplification of the exons of *GALNT2* for sequencing proved to be difficult. Getting good sequence data for exon 1 was a particular challenge although once betaine was used to relax the secondary structure of the area it was possible to get excellent quality sequencing data across that exon. The sequencing that was completed was only exonic and the immediately adjacent intronic areas. Other potential regulatory areas in intronic areas were not examined, which, if rs7536663 is associated with IBD susceptibility, may be important in the *GALNT2* gene. An attempt was made to do an analysis of conservation across species (data not shown) for the gene in order to select the best intronic areas for sequencing. This concentrated particularly on intron 1, but the problems of defining the best regulatory regions over such a huge intronic area (about 25kbp) for the purposes of selected sequencing meant that large areas of the intron would have needed to have been sequenced, something that was not possible with the available finances. With the development of more cost-effective high-throughput sequencing, this may not be a problem in the future.

This chapter also demonstrates the difficulties that can be encountered with genotyping on automated systems. Our group were the first to have genotyping done on the Illumina® platform at the WTCRF, Edinburgh, which meant that knowledge of the pitfalls of this technology were limited. Uncovering the problems proved difficult, and it was unfortunate that rs2281719 appeared to have such a good association initially. The genotyping for this SNP had to be completely discounted, not only because a large number of the DNAs fell into the undetermined genotype on reclustering, but also because some of the genotyping was miscalled despite being clearly within the centre of a cluster, or gave a completely opposite genotype to the true genotype. The fact that one of the ASOs (allele specific oligonucleotide) for that SNP has 2 perfect matches in the human genome may have been the problem, as suggested by Dr Gibbs at Illumina®. The ASO, as described in the materials and methods chapter, hybridises to genomic DNA, with extension to the LSO (locus specific oligonucleotide) as the 'tag' for the location. The subsequent PCR product has a complementary sequence to the LSO on the Illumina® bead to which it anneals and causes fluorescence of the bead. It is possible that the alternative site ASO on the genome, despite not having an associated LSO, would also have produced a PCR product and provided noise in the interpretation of the real signal. One of the other issues affecting the quality of the genotyping across all the Illumina® SNPs was DNA quality and concentration, which appears to be far more critical for Illumina® genotyping than for Taqman®. This required careful calling at an individual plate level to ensure accuracy.

Whether or not there is an association with the *GALNT2* gene and CD susceptibility, the novel interaction between *GALNT2* and *NOD2* suggested by the yeast two-hybrid screen could still be a biologically important interaction in the complex biological pathways of IBD. This interaction is investigated further in the next chapter.

Chapter 6 NOD2 and GALNT2 expression and interaction

Summary

Aims: To test the hypothesis that GALNT2 and NOD2 interact in mammalian calls and that they are expressed in the same cell types in the gastrointestinal tract.

Methods: Co-immunoprecipitation and western blotting in NOD2-transfected SW480 cells was used to investigate the interaction between NOD2 and GALNT2. Truncated NOD2 and the common NOD2 mutants were also transfected to investigate the position on the NOD2 protein with which GALNT2 interacted, and how the common NOD2 mutations affected the interaction. Immunohistochemistry was used to investigate whether GALNT2 and NOD2 were co-expressed in gastrointestinal tissue biopsies. Quantitative PCR was used to investigate how NOD2 and GALNT2 stimulators affected NOD2 and GALNT2 expression.

Results: NOD2 and GALNT2 were shown to interact in a mammalian system, with an interaction at the level of the CARD domain of NOD2. The common NOD2 variants reduced the intensity of the interaction. GALNT2 was shown to be expressed in enterocytes, goblet cells and the lamina propria, with the suggestion of a reduced expression in inflamed tissues. Due to the lack of a functioning NOD2 antibody, despite attempts to generate one, it was not possible to show that NOD2 was co-expressed. Quantitative RT-PCR did not show that NOD2 or GALNT2 stimulators changed GALNT2 or NOD2 expression respectively.

Conclusions: It has been proven that NOD2 and GALNT2 interact in a mammalian system, validating the yeast two-hybrid experiment that initially suggested the interaction. Further work is required to investigate whether NOD2 and GALNT2 are expressed in the same cell types and cellular location. Key to this is the development of a functioning NOD2 antibody.

6.1 Introduction

In Chapter 1 the yeast-two hybrid experiment and the interaction of GALNT2 with NOD2 in yeast cells was described. Experiments examining germline variation in GALNT2 and association with IBD have also been described in Chapter 5. The aim of this chapter was to test the hypothesis that GALNT2 and NOD2 interact in mammalian calls and that they are expressed in the same cell types in the gastrointestinal tract.

6.2 Validation and preparation of reagents specific to the studies in this chapter

Details of the standard cell culture, gel electrophoresis, western blotting, immunohistochemistry and PCR techniques are given in Chapter 2.

6.2.1 Validation of the commercial GALNT2 antibody

Protein lysate from SW480 cells was used in gel electrophoresis and blotted onto PVDF membrane. The membrane was probed overnight with a 1:500 dilution of GALNT2 antibody (final concentration 500pg/µl, Sigma-Aldrich); the secondary antibody was goat anti-rabbit IgG-HRP (400µg/µl, SantaCruz Biotech, Santa Cruz, CA). A single clean band of just over 60kDa, appropriate to the size of the GALNT2 protein (64.4kDa) was obtained (Figure 6-1).

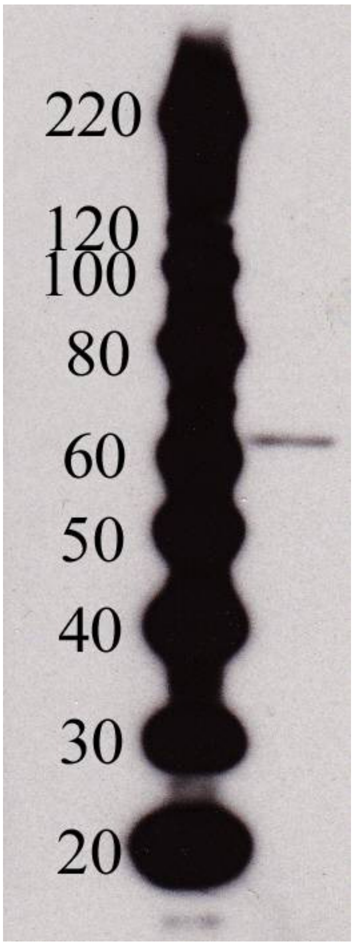


Figure 6-1 GALNT2 antibody Western blot with Magic Mark markings in kDa

6.2.2 Validation of commercially available NOD2 antibody

The most widely available NOD2 antibody in the literature is a mouse monoclonal antibody (clone 2D9) developed against the N-terminal end of the recombinant NOD2 protein (Cayman Chemicals, Ann Arbor, MI); this was used in a study demonstrating NOD2 protein in Paneth cells in the GI tract[264] and showed clear staining in Paneth cells of the terminal ileum.

Immunohistochemistry (IHC) was performed using this antibody. Unfortunately, despite repeated attempts with varying conditions (citrate buffer pH 6.0 antigen retrieval, protein blocking, and varying concentrations of antibody), there was no apparent staining of Paneth cells, or of any other cells in the terminal ileum.

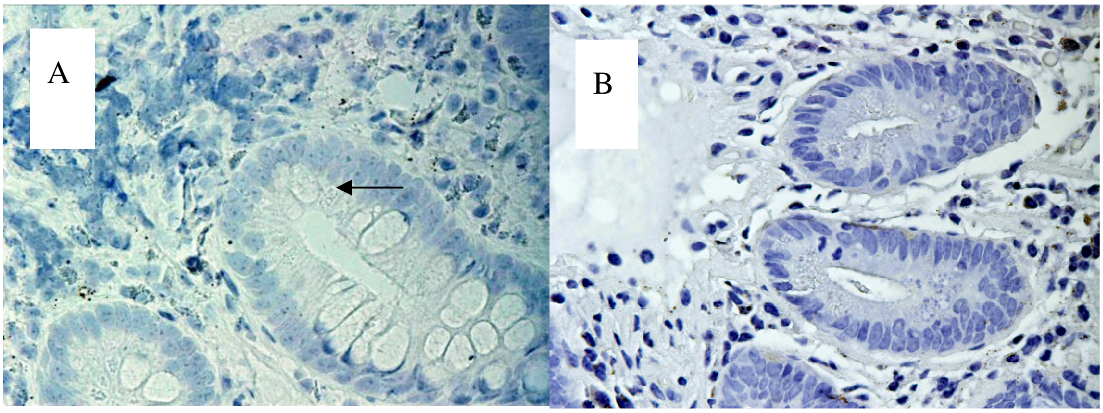


Figure 6-2 Terminal ileal biopsies using NOD2 Cayman Chemicals antibody, and stained with haemotoxylin. A=CD patient with mild inflammatory cell infiltrate, arrow points to Paneth cell; B=CD patient with normal terminal ileum

To determine whether the antibody recognised the denatured NOD2 protein, a western blot was completed by Dr Craig Stevens, using HCT116 cells that had transfected (or mock transfected) with an HA-tagged NOD2wt plasmid. The antibody successfully detected HA-tagged NOD2, but not endogenous NOD2 (Figure 6-3).

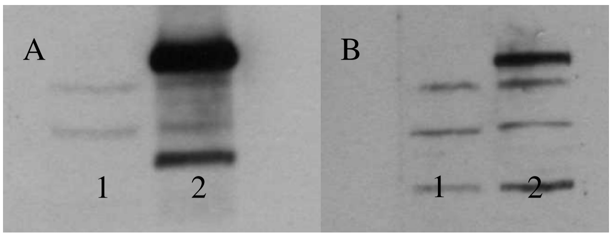


Figure 6-3 Western blot of mock transfected (1) and HA-tagged NOD2 wt transfected (2) HCT116 cells probed with HA antibody (A) and Cayman Chemicals NOD2 antibody (B)

6.2.3 Generation of a polyclonal NOD2 antibody

An extensive search was made for other commercially available antibodies; however all of them appeared to have been generated from the 2D9 clone used above. As the NOD2 protein is not commercially available, any attempt at antibody generation needed to be on the basis of selecting and generating a peptide sequence in a larger immunogenic vector for injection into an animal.

6.2.3.1 Peptide sequence selection

As the antibody was needed for both IHC and western blotting it was important that the peptide sequence chosen was on the exposed region of the protein when normally folded: protein modelling was therefore required. Dr Dinesh Soares (Medical Genetics group, Molecular Medicine Centre, University of Edinburgh), a bioinformatician specialising in using computer modelling, performed the protein structure prediction. A hydropathy plot of full-length NOD2 predicted the secondary structure (Figure 6-4) and enabled the generated of an approximate 3D model (Figure 6-5).

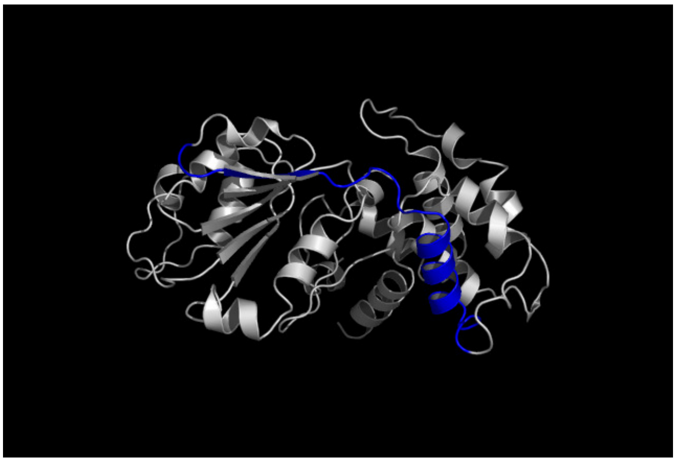


Figure 6-4 NOD2 NBD domain - predicted secondary structure

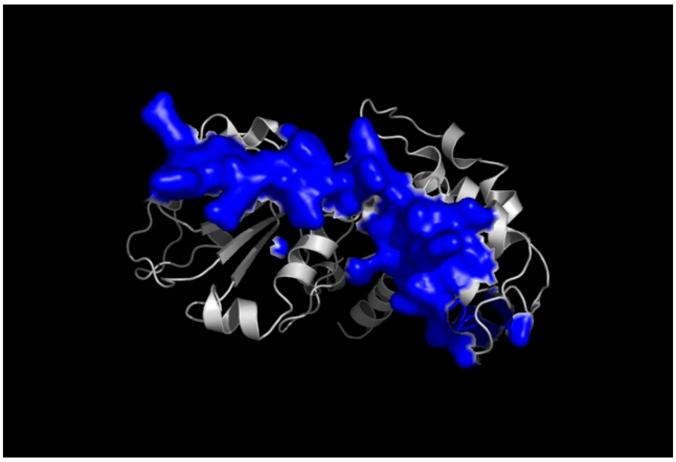


Figure 6-5 NOD2 NBD domain - predicted surface epitope

Several regions appeared to be suitable on the basis of their hydropathy scores and their exposure in the folded state of the protein. The amino acid sequence 435-454 was chosen as one of the best regions (sequence RKYIRTEFNLKGFSEQGIEL).

6.2.3.2 Antibody generation

The antibody generation was carried out in rabbits. To increase the immunogenicity of the peptide sequence, it was coupled to a large protein that would elicit a strong antibody response. The choice was between coupling it to keyhole limpet haemocynin (KLH) or manufacturing it as an octomeric multiple antigenic peptide (OMAP). The NOD2 peptide sequence was found to have 90% homology with rabbit NOD2. In view of this, it was thought that coupling to KLH would be more immunogenic and would increase the chance of successful generation of a NOD2 antibody. The antibody generation was provided by Alta Bioscience (University of Birmingham, UK) who coupled the NOD2 peptide sequence to the carrier protein and injected it into a rabbit at monthly intervals for 3 months. Three bleeds were undertaken at monthly intervals and the serum shipped to the GI laboratory, before a terminal bleed. The resulting serum from the terminal bleed was affinity purified by Alta Biosciences using controlled pore glass to separate the specific antibody from the rest of the proteins in the rabbit serum.

Each of the initial sera as well as the affinity purified terminal bleed serum were tested in western blots using protein lysates from SW480 cells that had been mock transfected, NOD2-transfected and left unstimulated, or NOD2 transfected with TNFα stimulation (to induce both endogenous and transfected NOD2).

None of the blots probed with the unpurified sera from the 3 initial bleeds (either at a dilution of 1:25 or used neat, data not shown) or the affinity-purified serum (Figure 6-6A) showed a band of the correct size. The single band that was picked up with the affinity-purified serum was 30kDa, far below the expected NOD2 band size of 115kDa. The blots were incubated with an antibody against the plasmid vector which gave a band of a different size to that seen with the anti-NOD2 antibody, indicating that the NOD2 antibody was not detecting the NOD2 protein (Figure 6-6B).

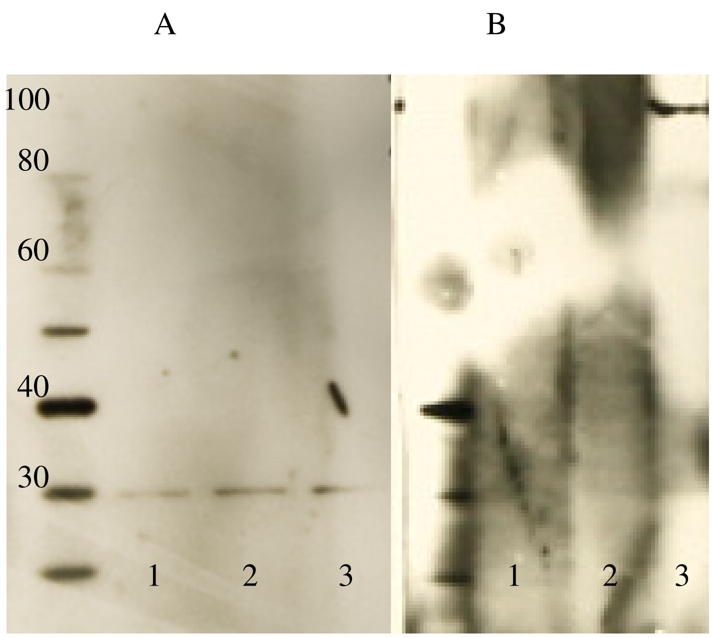


Figure 6-6 Western blots Probing with A: Alta generated NOD2 antibody 1:250; B: HA antibody 1:1000 1=Mock transfected cells, 2=Empty HA vector transfected, 3=NOD2 wt transfected. Magic Mark used as markers, numbers are in kDa

IHC was also performed with the Alta generated NOD2 antibody on ileal tissue. Three different sets of ileal tissue were used; all were from patients with UC and were reported to be within normal limits (i.e. without inflammation) when formally reported by consultant pathologists at the Western General Hospital, Edinburgh. Negative controls using PBS instead of antibody were completed as well as incubations with the antibody at 1:100 and 1:250 dilutions. All samples were subject to antigen retrieval using a pH 6.0 Citrate buffer, as already described.

There was a large amount of non-specific staining of the lamina propria and no Paneth cell staining in all three sample sets, as shown in Figure 6-7 (patient 1), Figure 6-8 (patient 2) and Figure 6-9 (patient 3).

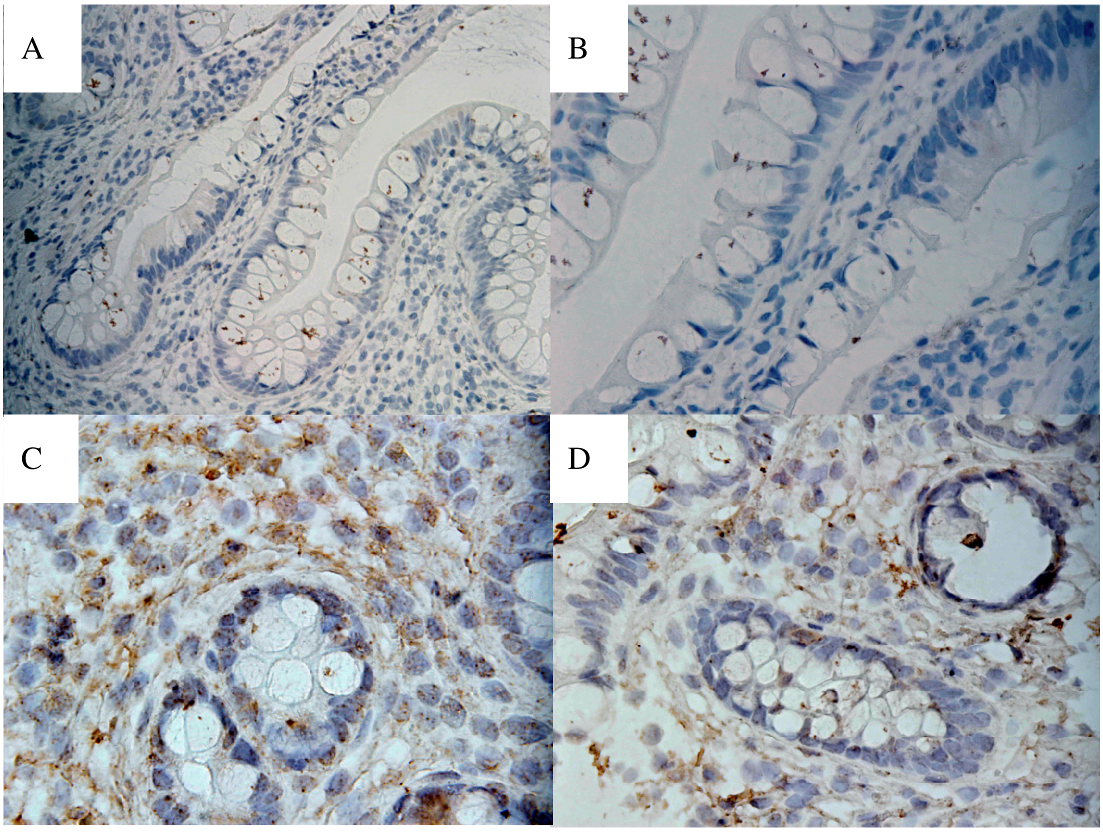


Figure 6-7 IHC: Patient 1; A&B=Negative controls, C=1:100 Alta antibody, D=1:250 Alta antibody

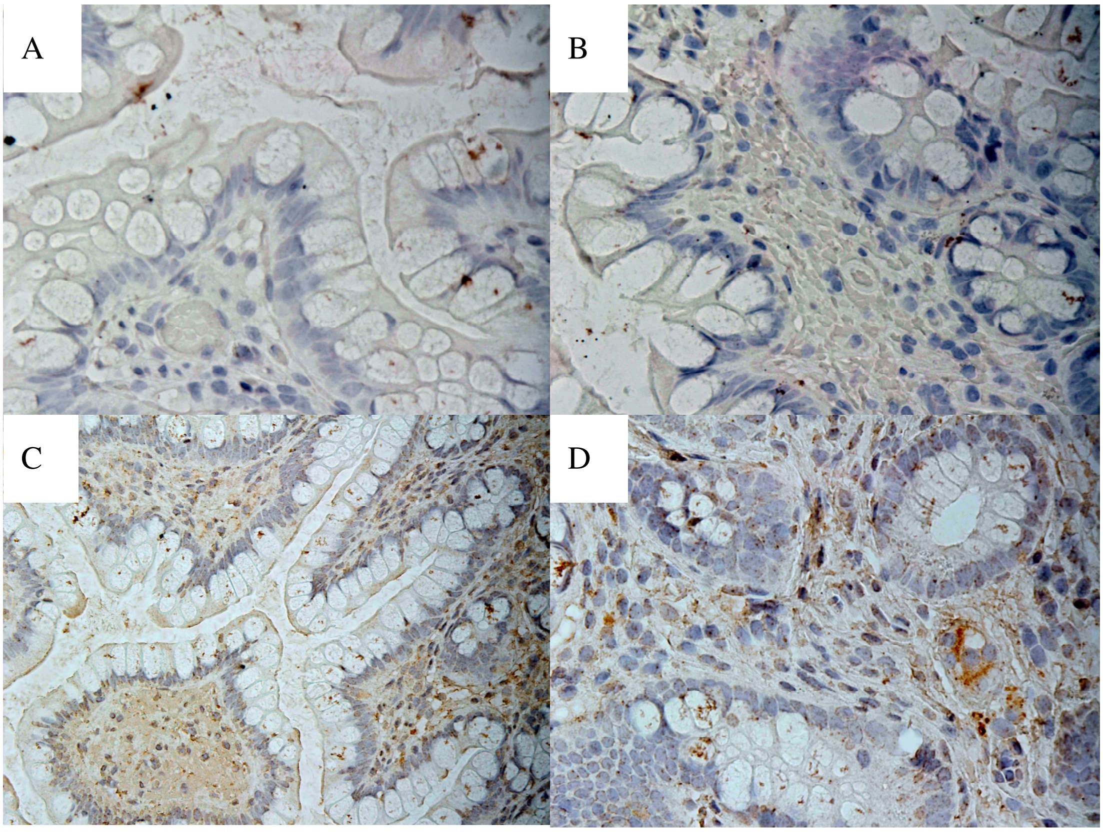


Figure 6-8 IHC: Patient 2; A&B=Negative controls, C&D=1:100 Alta antibody

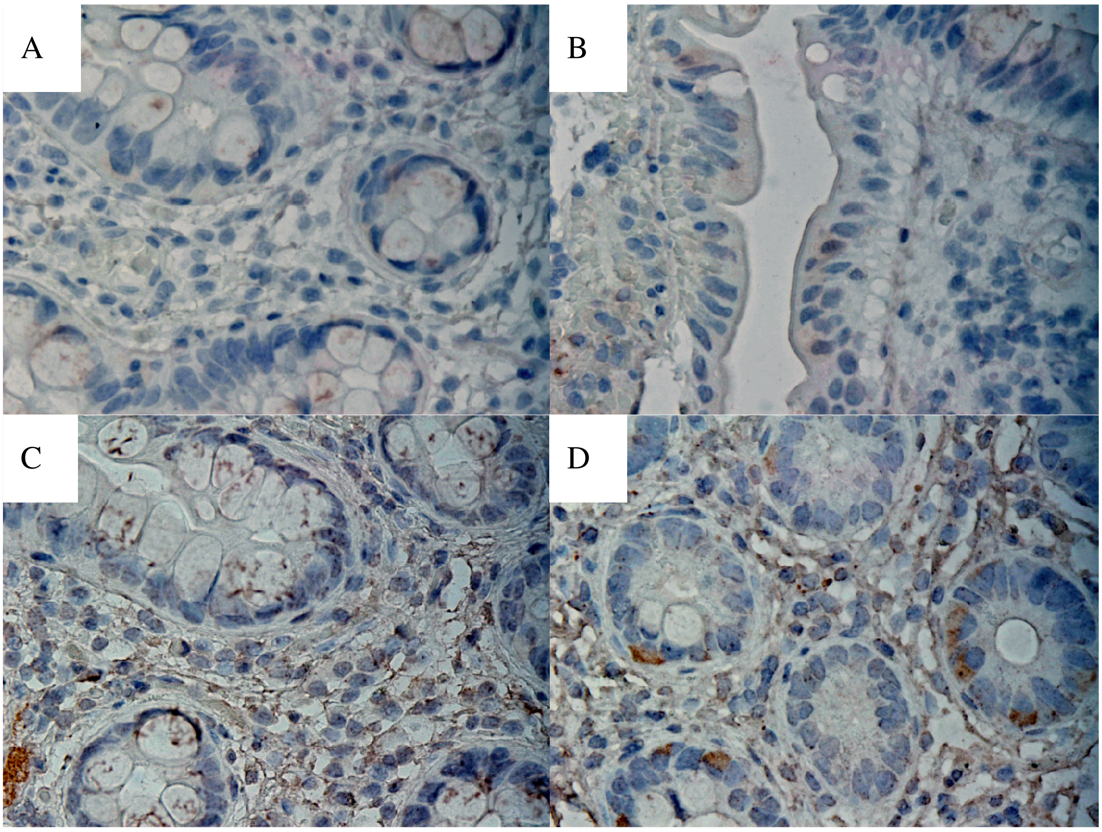


Figure 6-9 IHC: Patient 3; A&B=Negative controls, C=1:100 Alta antibody, D=1:250 Alta antibody

These experiments, along with the western blot of the purified antibody, indicated that it was unlikely that the antibody generation attempt had been successful.

Thus it was decided that obtaining a functioning NOD2 antibody for use in IHC and western blotting was beyond the scope of this research project. As an alternative approach for the co-immunoprecipitation (Co-IP) studies and western blotting, cells were used that had been transfected with tagged-NOD2 plasmids (containing wild type NOD2 or each of the three common variants) and antibodies to the plasmid tag were used for the western blot.

6.2.4 Site-directed mutagenesis – NOD2 G908R mutant production

As detailed in Chapter 2, the Quikchange® site-directed mutagenesis kit (Stratagene®, La Jolla, CA) was used to introduce the NOD2 G908R mutation into wild-type NOD2 that had previously been cloned into a pCMV-Myc vector (BD

Bioscience, gift of Dr Elaine Nimmo). Primers were designed by Dr Elaine Nimmo using the primer design guidelines given in the Quikchange manual; both primers covered the SNP of interest and contained the complementary base to the mutation rather than to the wild type. The primers chosen were 331/332: 5'CCTGGGATTCTGGCGCAACAGAGTGGG3' (forward), 5'CCCACTCTGTTGCGCCAGAATCCCAGG3' (reverse). In addition to a control reaction, two reactions with different quantities (2ng and 10ng) of the NOD2 wild type DNA template were used. The PCR reaction extension phase of each cycle was 3 minutes as the plasmid was approximate 3kbp in length. Following *Dpn* I restriction enzyme digest to remove the methylated strand, the PCR product was used to transform XL1-blue *E.coli* cells, with an overnight incubation at 37°C. The mutagenesis efficiency was calculated for the pWhitescript control plasmid. On this plate there were 418 blue colony forming units (cfu) and 66 white cfu, giving a mutagenesis efficiency of 86.4%. The transformation efficiency was also calculated for the pUC18 control plate (with 671 cfu) to be $>10^8$ cfu/µg.

To check whether the mutation had been introduced by the site directed mutagenesis, amplification was performed using primers 33/34: 5'AGGCCACTCTGGGATTGAG3' (forward), 5'GTGATCACCCAAGGCTTCAG3' (reverse) on DNA from 12 different colonies, 2 positive controls with the G908R mutation and 2 negative controls with wildtype NOD2. Sequencing of the PCR products was performed at the MRC Human Genetics Unit, Edinburgh. The results showed that 11 of the 12 colonies selected appeared to have been successfully transformed from the wild-type to the G908R mutation (Figure 6-10, A-K). The negative controls (N and O) were both wildtype; one of the positive controls was NOD2 G908R (M) and the other control sequence was too poor to interpret (not shown).

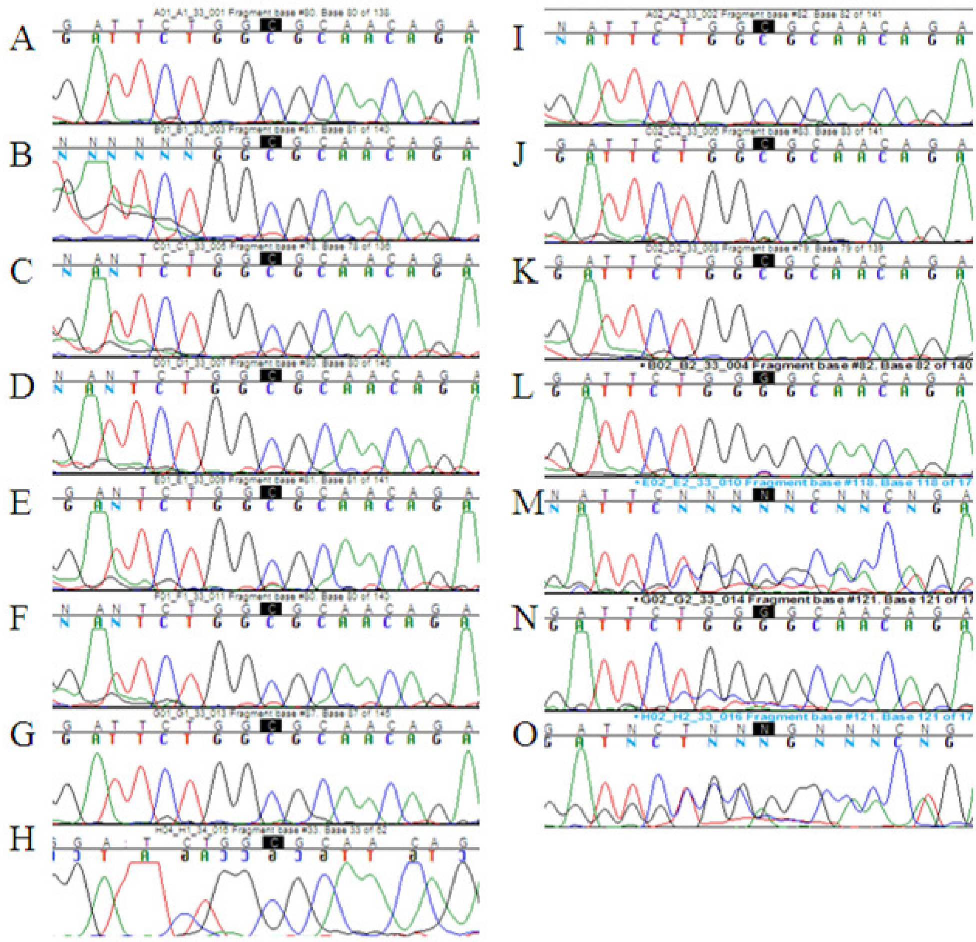


Figure 6-10 NOD2 G908R mutagenesis sequencing results, nucleotide in black is the 908 position A-K=successful mutagenesis, L=unsuccessful mutagenesis, M=positive control, N&O=negative control

Having confirmed that the site-directed mutagenesis had been successful, full length *NOD2* was sequenced in three of the colonies in which the mutation had been introduced to ensure that no other mutations had also been introduced. Three sets of primers were required to sequence across the full length *NOD2* (289/290, 291/32 and 292/293, a gift of Dr Elaine Nimmo). All primer sequences are given in Table 6-1. Gradient PCRs were set up to calculate the optimal annealing temperatures for each primer pair in the subsequent PCRs (Table 6-1). PCRs were set up with DNA isolated from the three colonies and sequencing of the PCR products was done at the MRC Human Genetics Unit. The primer combination 289/290 gave a PCR product of the wrong size (400bp rather than 840bp). The gradient PCR for these was repeated with double the normal amount of Taq (working concentration 1U/ml rather than 0.5U/ml), but the band size remained incorrect.

As an alternative approach, two pCMV-vector primers were used that were near the two junction points with the cloned *NOD2* gene (primers 398/399) and therefore amplified across the whole of the cloned *NOD2* gene. The optimal temperature for the annealing phase was 58°C, and as the PCR product was large (approximately 4000bp), a 4 minute extension phase was used. This amplification was performed with DNA from the 3 colonies and the PCR product sent for sequencing with a number of primers (as given in Table 6-1) to ensure that the entire gene was sequenced. Figure 6-11 gives an overview of the sequencing strategy and is shown with the *NOD2* exonic reference sequence. From examination of the DNA sequence obtained from all 3 colonies it was determined that apart from the G908R mutations no other mutations had been introduced.

PCR primer set	Forward primer 5'-3'	Reverse primer 5'-3'	Annealing temperature/°C
289/290	CCTAATGGCATATGATGGGGGAAGAGGGTGGTTCAGCC	CATGTGCCATGGGTGGCCAGGGGTGCTGAAGAGCTCC	N/A
291/32	CCCATGGCCATGGCCACATGCAAGAAGTATATGG	GGATGGAGTGGAAGTGCTTG	61
292/293	GCGCCCCTGGAATTCCTTCACATCAC	CCGGGATCCTCAAAGCAAGAGTCTGGTGTCCCTG	64
398/399, used with:	GATCCGGTACTAGAGGAACTGAAAAAC	TCATCAATGTATCTTATCATGTCTGG	58 (4 min extn)
63		TGGAAGTCTTGCCCTGCAG	
109	ATGGGGGAAGAGGGTGGTTC		
111	GCCACATGCAAGAAGTATATGG		
112	GAATTACCAGTCCCATTGGC		
115		GCCTCCTGACAATGGCTG	
118	ATGTGCTCGCAGGAGGCTTTTCAGGCA		
289	CCTAATGGCATATGATGGGGGAAGAGGGTGGTTCAGCC		
290		CATGTGCCATGGGTGGCCAGGGGTGCTGAAGAGCTCC	
399		TCATCAATGTATCTTATCATGTCTGG	
842	GTCCTGTTAACCTTTGATGGC		
843		TCAGCAGGTACATATCTGTAGTGG	
844	CAGATCACAGCAGCCTTCC		
845		CTCGTCACCCACTCTGTTG	
35/846	GGCAGAAGCCCTCCTGCAGGGCC	TGAATGCAATTGTTGTTGTTAAC	62

Table 6-1 Primer sequences for NOD2 PCRs and sequencing

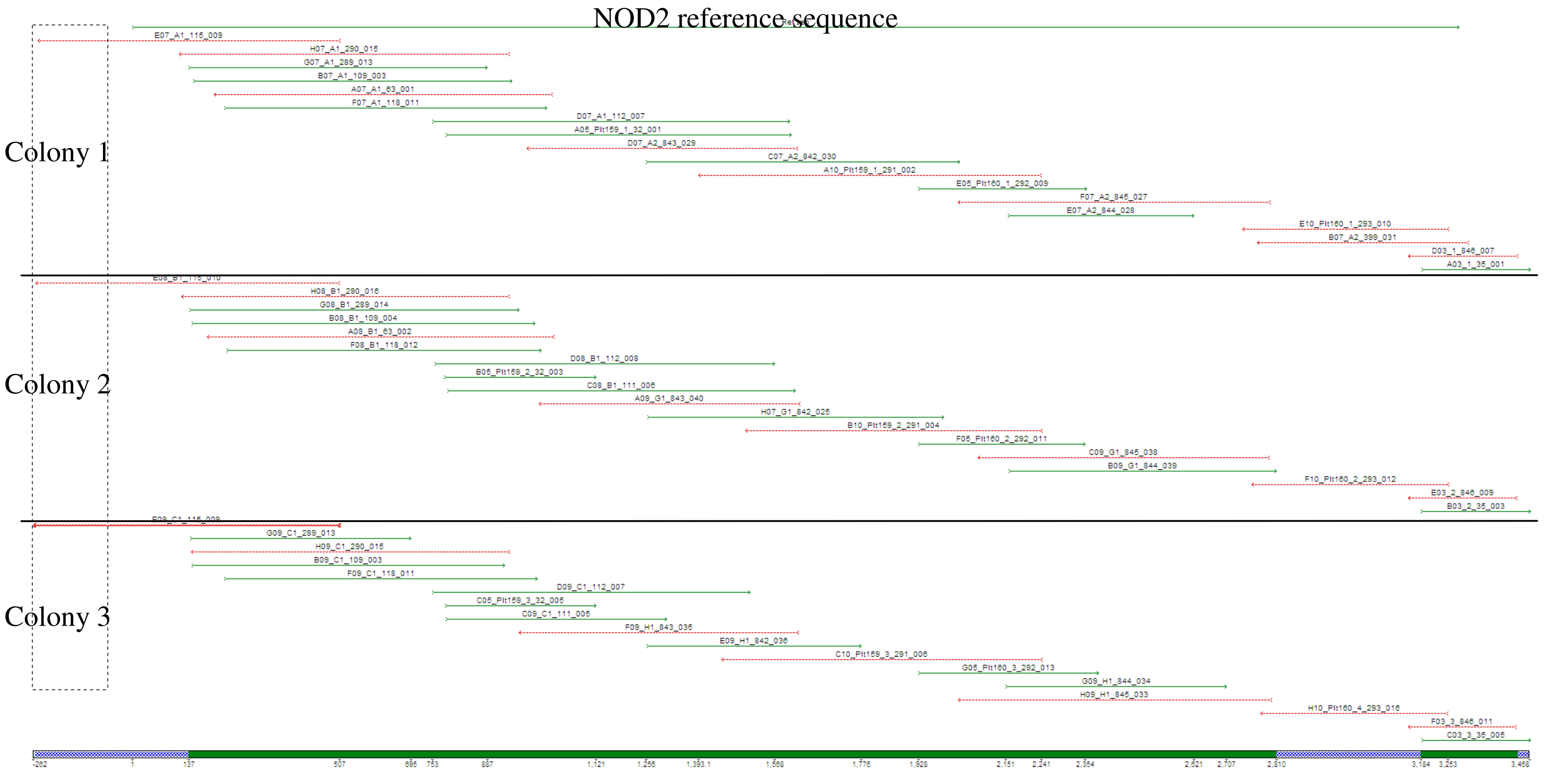


Figure 6-11 NOD2 sequencing for 3 selected colonies, indicating which primers were used in the sequencing. The NOD2 reference sequence is shown at the top of the figure, and is the entire exonic NOD2 sequence, without the 3'UTR region.

6.3 Co-immunoprecipitation experiments – NOD2/GALNT2 interaction in mammalian cells

The SW480 cell line was used in the co-immunoprecipitation (CoIP) experiments. This cell line was chosen because it had been used to generate the cDNA library for the initial yeast-two hybrid screen. Because endogenous NOD2 expression is low in SW480 cells, the cells were initially transfected with NOD2 cloned into the HA vector (gift of Dr Craig Stevens). The transfection was completed by transfecting 5µg of the appropriate DNA in Opti-MEM and lipofectamine into T25 flasks of SW480 cells at 80-90% confluence, and the cells incubated in Opti-MEM for 6 hours at 37°C before removing the complexes and medium and replacing it with complete medium. Cell lysate was prepared and protein levels quantified. In the Co-IP, GALNT2 was pulled down with the anti-mouse GALNT2 antibody (Sigma-Aldrich). As the antibody was relatively dilute and expensive, only 0.25µg was used rather than 2µg (the usual amount used in the GI laboratory). After the Co-IP reaction had been completed, gel electrophoresis and western blotting was performed, using the Co-IP products. To demonstrate that the initial transfections had been successful, 20µg of each of the corresponding protein lysates were run in a 2nd gel. The resulting membranes were probed with an antibody against HA (HA.11 Clone 16B12, Covance, Princeton, NJ) in view of the lack of a functioning NOD2 antibody.

The NOD2 transfection DNAs used in this section were: NOD2 wild type, NOD2 1-247 (ie CARD domains of NOD2 only), NOD2 1-693 (ie CARD and NBD domains of NOD2 only), NOD2 R702W mutant, NOD2 G908R mutant and NOD2 1007fs mutant. All of these DNAs had been cloned into the pCMV-HA expression vector (BD Bioscience) and were gifts of Dr Craig Stevens, although, as described in section 6.2.4, the NOD2 G908R had been initially cloned into the pCMV-Myc expression vector (BD Bioscience). It was subsequently transferred into the HA expression vector by Dr Craig Stevens.

6.3.1 Interaction of NOD2 wild type and GALNT2

The aim of this CoIP was to show that NOD2 and GALNT2 interact in mammalian cells and to discover which part of the NOD2 protein interacts with GALNT2. NOD2 wild type, NOD2 1-693, NOD2 1-247 and empty vector plasmids (all HA-tagged) were transfected into the SW480 cells with a mock-transfected control (identical conditions except that PBS used instead of a plasmid). The western blots of the whole cell protein lysates of the transfected cells (A) and pulled down with GALNT2 antibody (B) are shown in Figure 6-12. The molecular weights of the products were: NOD2 wild type: 115.3kDa, NOD2 1-693: 77.2kDa and NOD2 1-247: 27.8kDa. Proteins of the correct sizes were seen from the transfected cells and the Co-IPs. This demonstrates that GALNT2 and NOD2 interact in mammalian cells; moreover, there is an interaction with all lengths of the NOD2 protein. This suggests that GALNT2 and NOD2 interact at the CARD domains, which lie at the N-terminal end of the protein and were present in all 3 of the shortened NOD2 forms.

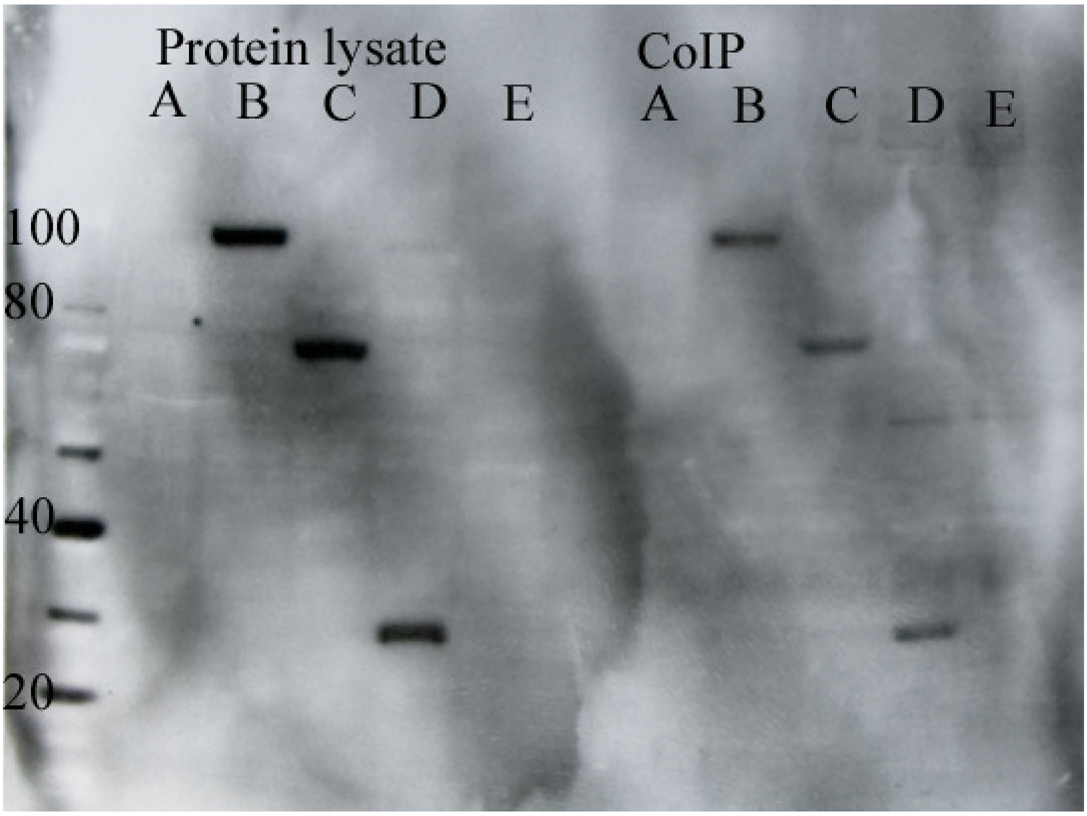


Figure 6-12 Western blot from CoIP pulling down GALNT2 and probing with HA antibody A=Mock transfection, B=NOD2 wild type, C=NOD2 1-693, D=NOD2 1-247, E=Empty vector

6.3.2 Interaction of variant NOD2 and GALNT2

The aim of these experiments was to investigate whether the NOD2 mutations affected the interaction with GALNT2. NOD2 wild type, NOD2 R702W, NOD2 G908R and NOD2 1007fs and an empty vector (all HA-tagged) were transfected into the SW480 cells. A mock transfection was also completed, as before. The expected molecular weights of the NOD2 wild type, NOD2 R702W and NOD2 G908R were 115.3kDa, whereas for the NOD2 1007fs it was 111.4kDa. The results are shown in Figure 6-13 (whole cell protein lysates) and Figure 6-14 (pulled down with GALNT2 antibody). The western blot of proteins in lysates of the transfected cells (Figure 6-13) demonstrates that the transfections were successful. The Co-IP (Figure 6-14) shows that GALNT2 interacted with each of the NOD2 variants, but the strength of the interaction varied, as suggested by the intensity of the band. The strength of the interaction was less with all the NOD2 variants compared with the wild type, but especially with the NOD2 R702W variant.

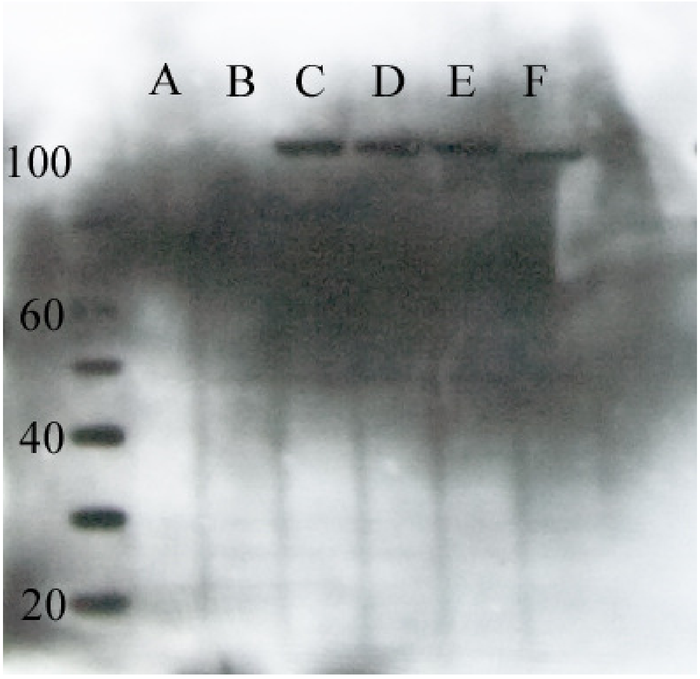


Figure 6-13 Western blot of protein lysates probing with HA antibody A=Mock transfection, B=Empty vector, C=NOD2 wild type, D=NOD2 R702W, E=NOD2 G908R, F=NOD2 1007fs

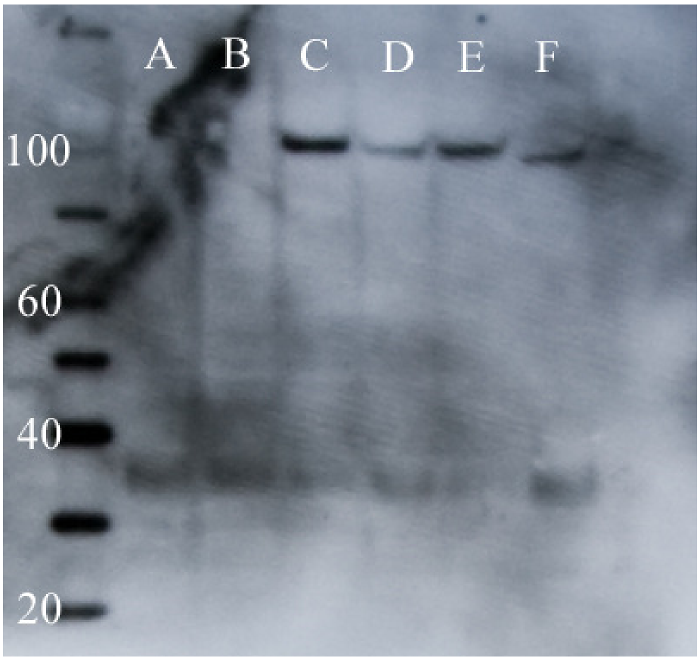


Figure 6-14 Western blot from CoIP pulling down GALNT2 and probing with HA antibody A=Mock transfection, B=Empty vector, C=NOD2 wild type, D=NOD2 R702W, E=NOD2 G908R, F=NOD2 1007fs

6.4 GALNT2 protein expression in gut biopsies

6.4.1 Methods

IHC was carried out on intestinal biopsy samples from patients with IBD using the same GALNT2 antibody as used for western blotting according to the protocols detailed in section 2.7.3. For this antibody, information was available on www.proteinatlas.org, including optimal concentrations and pictures of staining in normal and cancerous tissues. Initial experiments defined optimal conditions as heat induced epitope retrieval (HIER) using a pH6.0 citrate buffer and an antibody dilution of 1:50 (0.025µg/ml). Colonoscopic biopsies were used from patients with CD (n=7) and UC (n=2).

6.4.2 Results

In the uninflamed terminal ileal samples, there was some GALNT2 staining in the enterocytes and goblet cells, as well as in the lamina propria, as shown in Figure 6-15. GALNT2 appeared to be somewhat reduced in tissues that also had some inflammatory infiltrate (Figure 6-16). There was no evidence of GALNT2 staining of Paneth cells in any of the sections examined. In the colon, GALNT2 staining was predominantly in the lamina propria, with some on the enterocytes (Figure 6-17). In the few sections that had evidence of inflammation, antibody staining for GALNT2 appeared to be decreased (Figure 6-18).

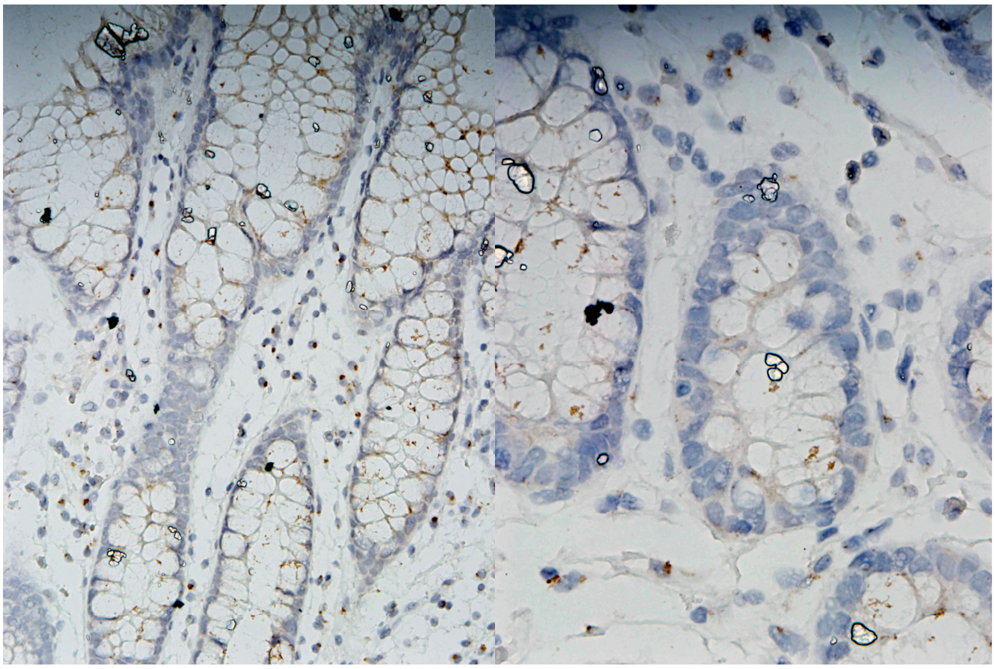


Figure 6-15 GALNT2 IHC of uninflamed terminal ileal tissue

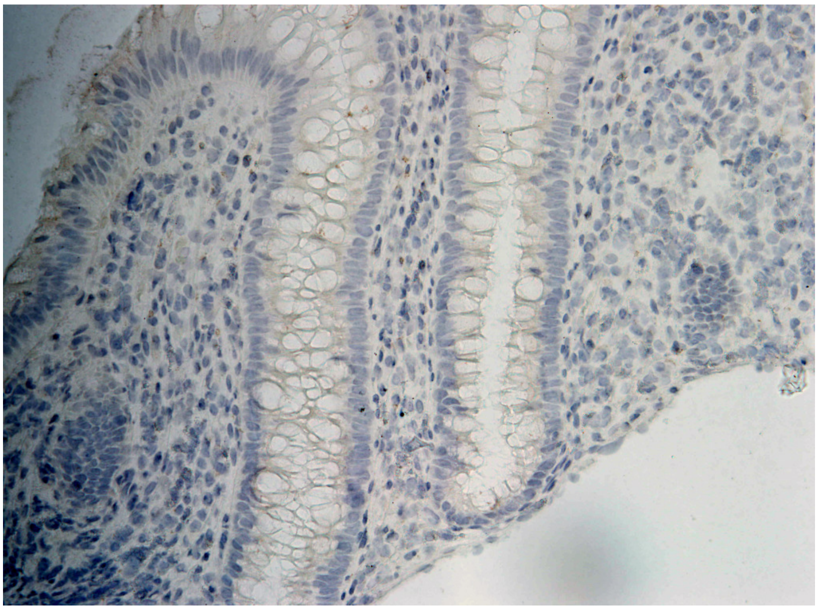


Figure 6-16 GALNT2 IHC of terminal ileal tissue having a patchy increase in chronic inflammatory cells.

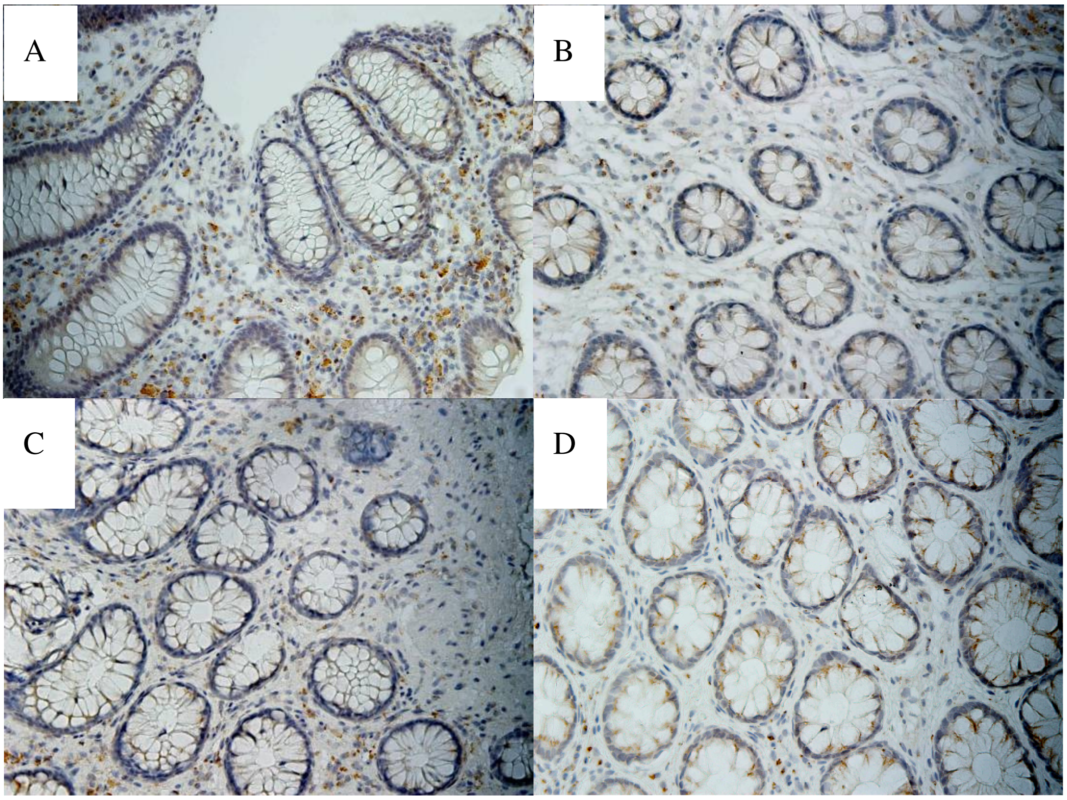


Figure 6-17 GALNT2 immunohistochemistry A=Ascending colon, B=Transverse colon, C=Descending colon, D=Sigmoid colon. All uninflamed tissues

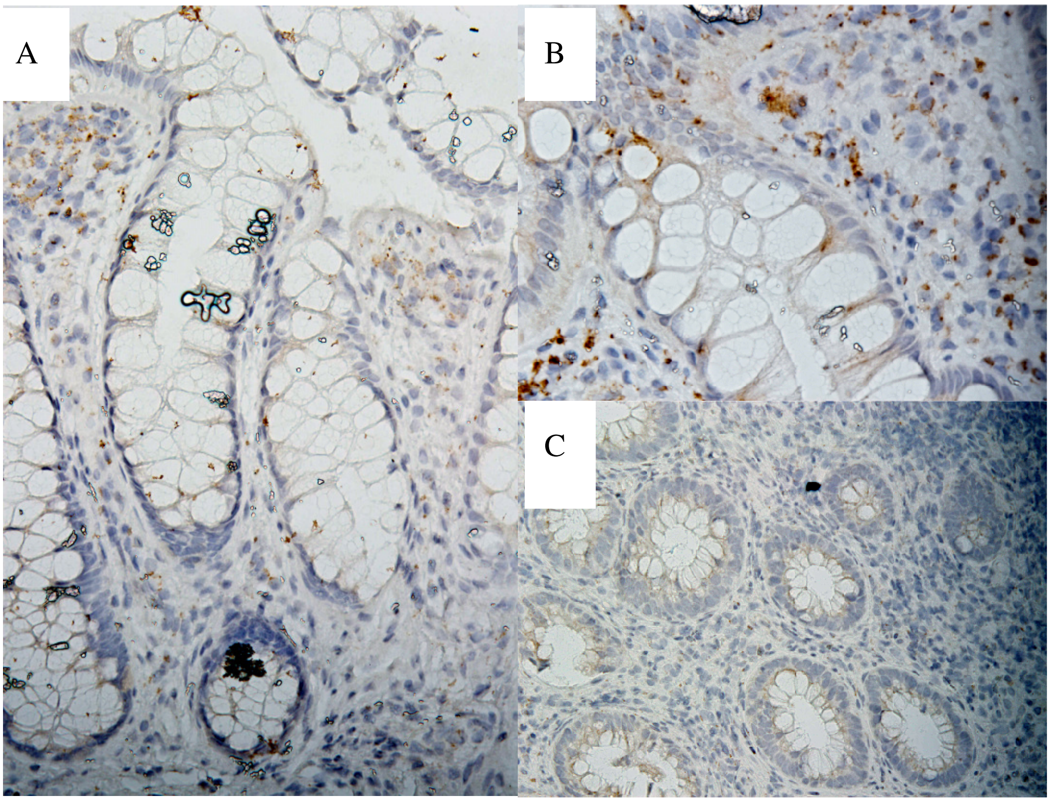


Figure 6-18 GALNT2 immunohistochemistry A&B=uninflamed rectum, C=rectum reported as having a mild patchy increase in inflammatory cells

6.5 Messenger RNA expression studies with GALNT2 and NOD2

GALNT2 and NOD2 expression was investigated using LS174T cells, a mucinous colon cancer cell line, which was thought to be most appropriate for investigating potential interactions between an enzyme involved in O-glycosylation and NOD2.

6.5.1 Methods

The primers used for qRT-PCR are listed in Table 6-2.

	Forward primer 5'-3'	Reverse primer 5'-3'
GALNT2	ATGGGCCTTGACGAAGGAGAAG	CCTCGATCTGTTCCCATTTCTGTC
GAPDH	TCATCTCTGCCCCCTCTGCT	CGACGCCTGCTTCACCACCT
NOD2 Taqman®	Taqman® inventoried assay: Hs00223394_m1 (ABI Biosystems)	
GAPDH Taqman®	Taqman® inventoried assay: Hs99999905_m1 (ABI Biosystems)	

Table 6-2 qPCR primers

6.5.2 PCR Optimization

6.5.2.1 SYBRGreen®

For each set of primers, amplification of serial dilutions of LS174T cDNA by PCR was set up as described in the Chapter 2. A standard curve was produced and concentrations of cDNA within the exponential doubling phase of the qPCR were used for subsequent qPCRs. From these experiments, it was determined that the best cDNA dilution was 1:100 for GALNT2 and for GAPDH.

6.5.2.2 Taqman®

The NOD2 Taqman® optimization PCR was run with several different sets of samples. A NOD2 standard curve used dilutions of a pCMV-Myc plasmid containing NOD2 wt (gift of Dr Elaine Nimmo). In addition, two standard curves of cDNA from two samples of LS174T cells were completed, one of which had been transfected with NOD2. The qPCR with the non-transfected cells showed extremely low levels of *NOD2* expression compared with the NOD2 plasmid controls and the

Ct values did not change in the expected way with serial dilution, as shown in Table 6-3. In the non-NOD2 transfected form, *NOD2* mRNA expression levels were possibly too low to quantitate accurately by qPCR; therefore *NOD2* expression was only examined in the time courses where NOD2 had been transfected. In the NOD2 transfected cells the optimal cDNA concentration for NOD2 qPCR was found to be 1:500, and for the Taqman® GAPDH qPCR 1:100.

cDNA dilution	Ct	Given Conc (copies/ul)	Calc Conc (copies/ul)	% Var
1:1	31.13	500	194710	38842%
1:1	31.03	500	201155	40131%
1:5	33.4	100	897.49	797.50%
1:5	33.34	100	1,442.78	1342.80%
1:10	33.96	50	20.06	59.90%
1:10	33.69	50	127.95	155.90%
1:50	34.43	10	0.81	91.90%
1:50	34.22	10	3.28	67.20%
1:100	34.41	5	0.92	81.60%
1:100	33.97	5	18.28	265.60%
1:500	34.39	1	1.04	4.10%
1:500	34.14	1	5.63	462.70%
1:1000	34.88	0.5	0.04	92.80%
1:1000	34.79	0.5	0.07	87.00%
1:5000	34.91	0.1	0.03	71.00%
1:5000	34.8	0.1	0.06	37.90%
1:10000	34.98	0.05	0.02	63.40%
1:10000	35	0.05	0.02	68.50%
1:50000	34.62	0.01	0.22	2085.00%
1:50000	34.66	0.01	0.17	1570.50%
0	35.38		0	
0	35.9		0	

Table 6-3 NOD2 Taqman® qPCR, non-transfected cells

6.5.3 Choice of stimulators

In the absence of a known positive control for up- or down-regulating *GALNT2* expression, monensin was used. Monensin is a carboxylic ionophore which acts as a Na^+/H^+ antiporter modifying Golgi pH and causing osmotic swelling of the Golgi cisternae, and thus inhibiting Golgi function.[265;266] As it would appear that cis-Golgi, where GALNT2 is thought to act[267], is affected later than trans Golgi[268], the 48 hour time point was examined. In published studies monensin has been used in varying concentrations; 1µM was used in the experiments in this chapter as this was the concentration used in a MUC2 study in LS180 cells.[269]

Carbachol, a cholinergic agonist which promotes mucin secretion from mucin producing cells, was used as a negative control as its promotion of mucin secretion is not through stimulation of the Golgi.[270] It was used at a concentration of 1mM, which was the concentration used in the MUC2 study in LS180 cells.[269]

In order to examine the effect of NOD2 on *GALNT2* expression, a range of NOD2 stimulators were chosen. TNFα is a cytokine that upregulates *NOD2* expression via NF-κB[271], and was used at a final concentration of 50ng/ml. Muramyl dipeptide (MDP), a peptidoglycan subunit which is a component of Gram positive and Gram negative bacteria, is a direct ligand of NOD2[89], and was used at a final concentration of 1µg/ml. Lipopolysaccharide (LPS), an outer wall component of Gram negative bacterial cells, is an endotoxin that stimulates NOD2 via the toll-like receptors[89], and was used at a final concentration of 1µg/ml.

Time points chosen for cell harvest were 0, 8, 24 and 48 hours. These were selected on the basis of previous experiments that showed TNFα-induced NOD2 expression in human epithelia cell lines[272] and mice macrophage cell lines[273] peaks at 6-8 and 24 hours.

Each set of time course experiments were triplicated for the untransfected time course and the NOD2wt transfected time course. An HA-empty vector transfection and a mock transfection time course were also completed. Each qPCR sample was run in duplicate.

6.5.4 Quality control

RNA samples were quantified by Nanodrop and quality checked by ensuring that the A260/280 was >1.8 and the A260/230 was >2.0. cDNA was made from standard amounts of RNA, as detailed in Chapter 2. Two negative template controls (NTC) were run for each qPCR run to ensure there was no DNA contamination in the primers or master mixes. A standard curve was run for each reaction to allow R^2 -

the correlation coefficient - to be calculated, as a measure of accuracy. The reaction efficiency was calculated for each run for both the target gene and the housekeeping gene in order to ensure they were similar and to allow normalization as detailed in the section 6.5.5.

6.5.5 Normalization

The time courses were normalized to glyceraldehyde-3-phosphate dehydrogenase (GAPDH) which had previously been established to be stable in relevant cell lines (Dr Marian Aldhous, unpublished data). As all the experiments were in the same cell line, variation in *GAPDH* expression levels between different cell populations was not an issue, so initial optimisation of qPCRs on a full panel of housekeeping genes was not required. As absolute quantification of copy number was not required, relative quantification was completed using the $2^{\Delta\Delta Ct}$ method[274], as the reaction efficiencies for *GAPDH* and the target enzyme were almost identical with both the SYBRGreen® and the Taqman® qPCR methods.

6.5.6 Statistical analysis

The two tailed independent t-test was used to compare each set of stimulations to the unstimulated set at each time point. A significant result was considered to be $P<0.05$.

6.5.7 Results: GALNT2 expression in non-transfected cells

Results were calculated as fold change from time point 0 and are shown in Figure 6-19 and Figure 6-20. No stimulator tested produced consistently statistically significant changes across the time points. Monensin reduced *GALNT2* expression compared with the unstimulated cells at 48 hours, but this did not quite reach statistical significance. The p-values for the comparisons are given in Table 6-4.

	Unstimulated vs TNFα	Unstimulated vs LPS	Unstimulated vs MDP	Unstimulated vs Monensin	Unstimulated vs Carbachol
8 hours	0.464	0.388	0.723	0.992	0.280
24 hours	0.599	0.879	0.907	0.293	0.696
48 hours	0.688	0.270	0.210	0.051	0.690

Table 6-4 GALNT2 expression Two-tailed independent t-test p-values for comparisons to the unstimulated time course

6.5.8 Results: GALNT2 expression in NOD2-transfected cells

For the NOD2 wild type transfection time courses, the graphs of results are shown in Figure 6-21 and Figure 6-22. The p-values for the comparisons are given in Table 6-5. There were no statistically significant differences in any of the comparisons.

	Unstimulated vs TNFα	Unstimulated vs LPS	Unstimulated vs MDP	Unstimulated vs Monensin	Unstimulated vs Carbachol
8 hours	0.953	0.618	0.650	0.649	0.887
24 hours	0.583	0.861	0.155	0.371	0.800
48 hours	0.571	0.678	0.453	0.133	0.713

Table 6-5 GALNT2 expression Two-tailed independent t-test p-values for comparisons to the unstimulated time course - NOD2 transfected time course

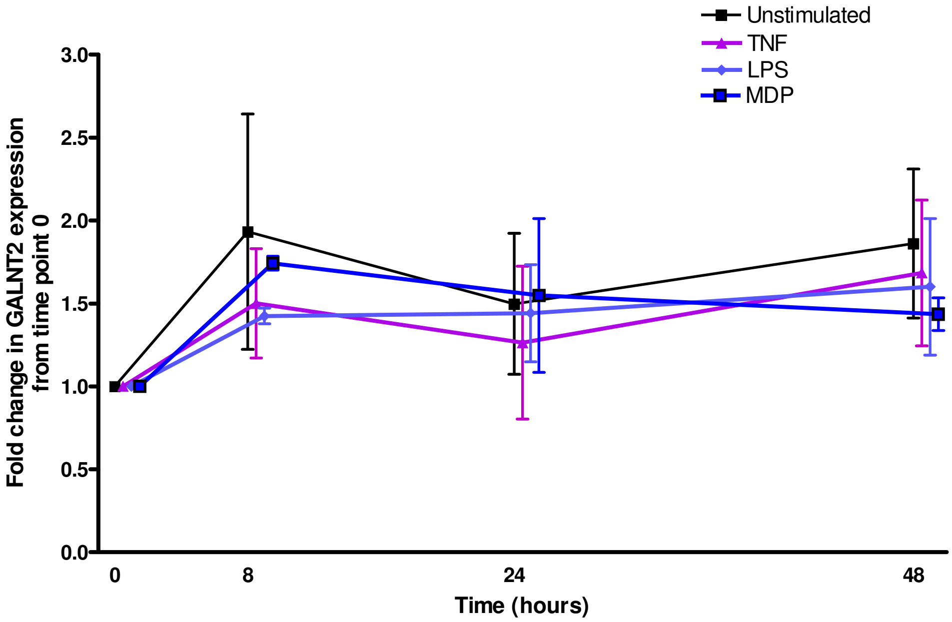


Figure 6-19 GALNT2 expression in unstimulated and TNF/LPS/MDP time courses. Data points are mean ± standard error of the mean (SEM)

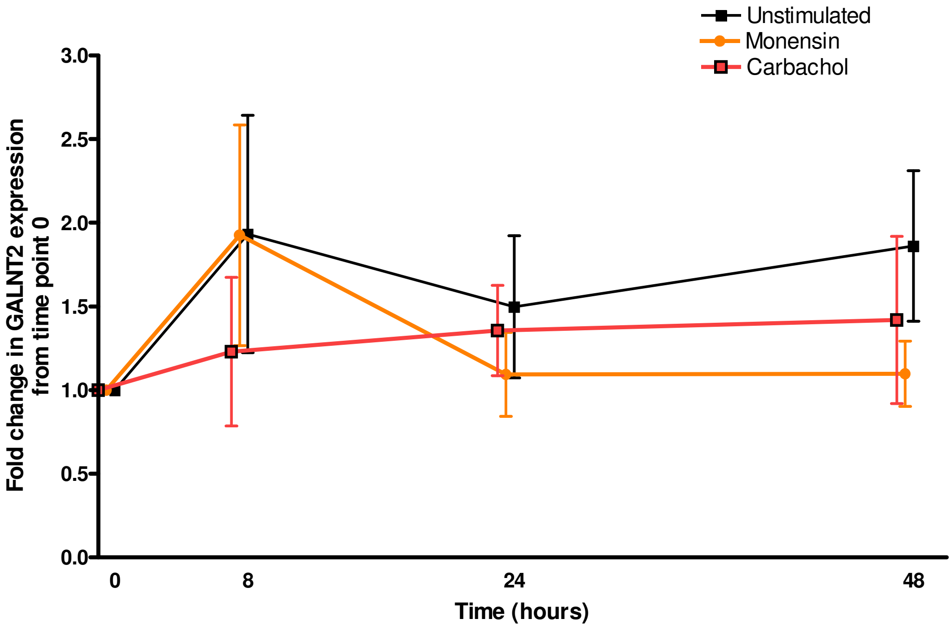


Figure 6-20 GALNT2 expression in unstimulated and monensin/carbachol stimulated time courses. Data points are mean ± SEM.

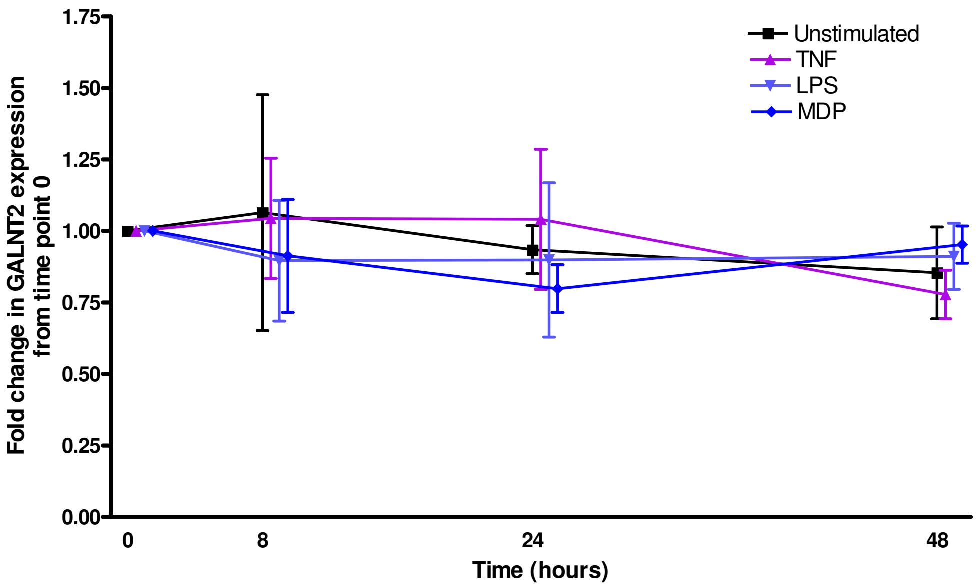


Figure 6-21 GALNT2 expression in NOD2 wild type transfected unstimulated and TNF/LPS/MDP time courses. Data points are mean ± SEM.

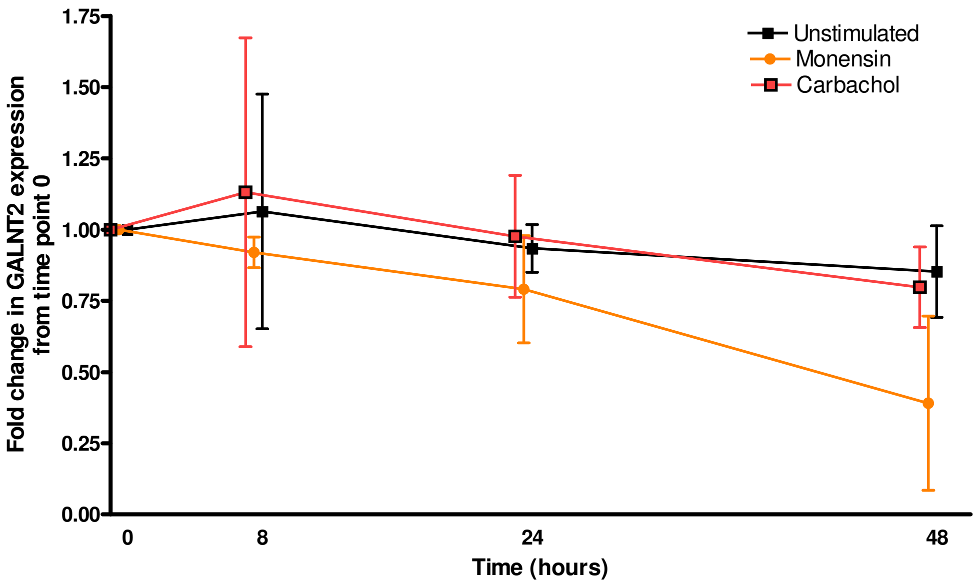


Figure 6-22 GALNT2 expression in NOD2 wild type transfected unstimulated and monensin/carbachol time courses. Data points are mean ± SEM.

6.5.9 Results: NOD2 expression in NOD2-transfected cells

Results from two time courses could be used, as the third experiment qPCR had one very poor R^2 value and differing reaction efficiencies between GAPDH and NOD2.

For the NOD2 wild type transfection time courses, the graphs of results are shown in Figure 6-23 and Figure 6-24. The p-values for the comparisons are given in Table 6-6. There were no statistically significant differences when compared to the unstimulated cells.

	Unstimulated vs TNFα	Unstimulated vs LPS	Unstimulated vs MDP	Unstimulated vs Monensin	Unstimulated vs Carbachol
8 hours	0.369	0.921	0.593	0.794	0.876
24 hours	0.417	0.213	0.178	0.133	0.205
48 hours	0.461	0.991	0.964	0.364	0.362

Table 6-6 NOD2 expression Two-tailed independent t test p-values for comparisons to the unstimulated time course - NOD2 transfected time course

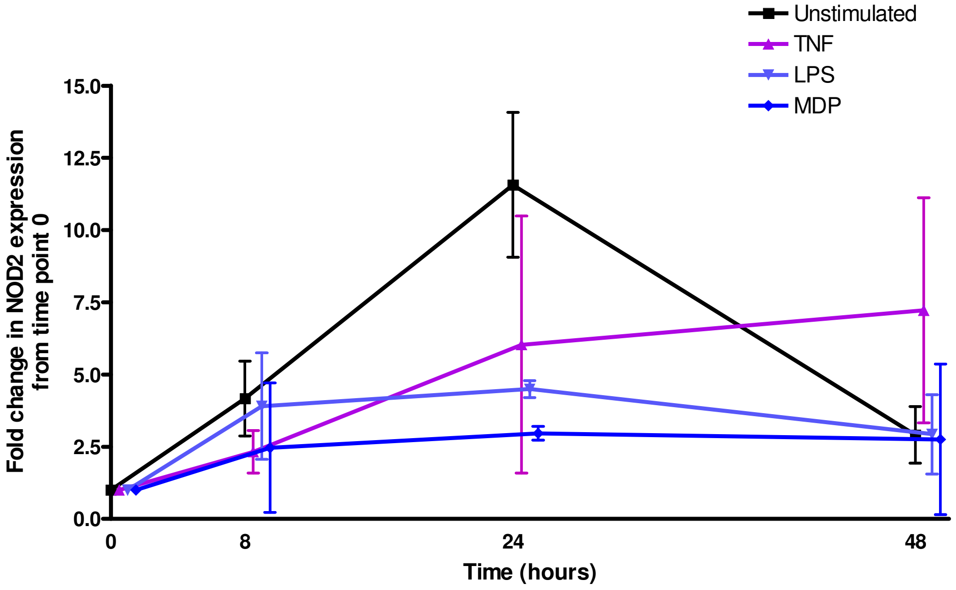


Figure 6-23 NOD2 expression in NOD2 wild type transfected unstimulated and TNF/LPS/MDP time courses. Data points are mean ± SEM.

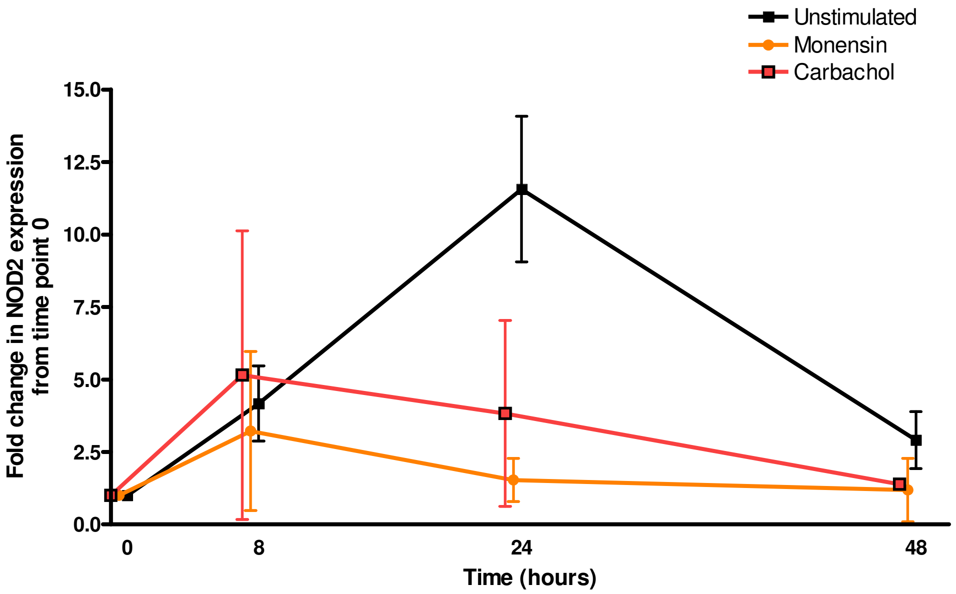


Figure 6-24 NOD2 expression in NOD2 wild type transfected unstimulated and monensin/carbachol time courses. Data points are mean ± SEM.

6.6 Discussion

This chapter has demonstrated convincingly that NOD2 and GALNT2 interact in mammalian cells. This validates the yeast two-hybrid experiment that showed an interaction in yeast cells. Interestingly, although GALNT2 appears to interact with the CARD domains of NOD2 (at the N terminal end of the protein), the interaction is affected by the common NOD2 mutations, which are all in the LRR domain of the protein (at the C-terminal end of the protein). The bands on the western blot were less intense from CoIP with the NOD2 variant compared with wild type suggesting that the interaction between NOD2 and GALNT2 is less strong with the variants. The R702W variant appeared to have the greatest effect on the interaction. As these experiments were completed once each, further experiments are required to confirm the results. In addition, densitometry on the western blot band intensities would be worth doing to quantitate the difference.

It is likely that the NOD2 mutants cause a conformational change in the tertiary structure of the NOD2 protein, meaning that the GALNT2-NOD2 interacting area of NOD2 is less exposed. Interestingly, the R702W mutant is a change from the hydrophilic, polar amino acid arginine to the non-polar tryptophan with a neutral hydropathy index. However, the G908R mutation is a change from the non-polar glycine (with a neutral hydropathy index) to arginine. Either of these changes could substantially change the tertiary structure of the protein, thus affecting the interaction of NOD2 with the other proteins and affecting downstream functions of NOD2.

An alternative cell line with higher endogenous levels of NOD2 expression could be used to explore the interaction of NOD2 and GALNT2 by Co-IP which would be useful in case the artificially high level of NOD2 in transfected cells affected the nature of the interaction.

Having shown that NOD2 and GALNT2 interact in mammalian cells, the next question was to address whether the two proteins are expressed in the same cellular compartments of the same cell types. NOD2 is a cytosolic protein whereas GALNT2 is located in the Golgi apparatus. Because a NOD2 antibody was not available either commercially or in-house, it has not been possible to prove that these proteins are expressed in the same cells. Certainly this chapter has shown that GALNT2 is expressed in enterocytes, goblet cells and the lamina propria, with a possible reduction in GALNT2 expression in inflamed tissues. The appropriate IHC with a functioning NOD2 antibody, when available, would be the next step. To investigate whether they are both expressed in the same cellular compartments, immunofluorescence microscopy or immuno-electron microscopy could also be helpful, again, only if a functioning NOD2 antibody was available. Co-IP after separating different cell fractions could also be an alternative method to investigate if NOD2 and GALNT2 are expressed in the same cellular compartments.

The commercially available NOD2 antibody did not stain appropriately in the IHC presented in this chapter, despite there being clear pictures in the literature of beautiful staining of Paneth cells. Why did it not work? The lack of an appropriate band on the western blot - unless NOD2 transfected protein lysate is used - would suggest that there is a problem with the NOD2 antibody itself, as long as the HCT116 cells used in this experiment produced sufficient of endogenous NOD2. It is possible that the endogenously produced NOD2 does not express the N-terminal end of the protein, thus, as the 2D9 clone is raised against the N-terminal end of the protein, it would not recognise the endogenous protein. It would have been helpful to have contacted the authors of the paper[264] which demonstrated the antibody staining Paneth cells in case there were further IHC conditions, not detailed in the paper, which may have worked.

The attempt to raise a NOD2 antibody was also unsuccessful as it did not generate a band of the correct size on the western blot and it also failed to show appropriate staining on IHC. It is likely that the initial antibody generation was unsuccessful. The fact that the amino acid sequence had 90% homology with rabbit should not have mattered, as it was coupled with KLH. However, only a tiny proportion of the antibodies generated might have been against NOD2, with the rest of the antibodies against peptide sequences on KLH. With the benefit of hindsight both methods of peptide conjugation should have been used to increase the chances of successful antibody production. OMAP is a method of peptide antigen synthesis where the peptide is synthesised directly onto a branching lysine core, negating the need for a carrier protein and potentially producing a higher proportion of antibodies against the target peptide sequence. Whether the 90% homology of the peptide with rabbit sequence - the initial concern that led to the choice of KLH - would have been an issue with OMAP is not clear.

The best way of successfully producing an antibody is to use the relevant purified protein as the immunogen, as this ensures that it is presented in its most biologically relevant state, with the outermost epitopes on display. However, the anti-bacterial properties of NOD2 make bacterial-based protein production methods difficult. If in the future it is possible to produce large amounts of NOD2 protein, this should be used in a further antibody generation attempt. The fact that there is not a good commercially available antibody available, despite the scientific need, would suggest that biotechnology companies have also been unsuccessful in their NOD2 antibody generation attempts.

Alternative ways needed to be found to examine for an interaction between GALNT2 and NOD2. Using qPCR with NOD2 and GALNT2 stimulators and investigating if mRNA levels of either increased was an appropriate alternative, with the caveat, of course, that proteins which affect the production of another protein do not necessarily interact: it could happen through a third protein. The data presented here do not show that NOD2 stimulators change *GALNT2* mRNA expression. It would be worthwhile repeating the experiments with cells demonstrating a higher endogenous expression of *NOD2* mRNA, as the low levels of *NOD2* expression in the cells may mean that the interaction is not relevant in LS174T cells. Ideally, a cell line with quantifiable *NOD2* levels should be used. It would also be interesting to use cell lines which are known to have germline variation in the *NOD2* gene, especially those with the 3 common *NOD2* variants.

Expression in cell lines is not necessarily an accurate reflection of what happens in vivo as cell lines are transformed and the cells could have different expression profiles compared with normal cells and be regulated differently. Therefore, looking at mRNA expression in patient biopsies would also be useful, although that would require a different housekeeping gene for normalisation. Many types of cells would be present which may introduce heterogeneity into the samples. A way round this would be to use microdissection prior to RNA extraction, although the problem would then be the small amounts of material available.

In summary, this chapter has demonstrated that NOD2 and GALNT2 interact in a mammalian system. Further work is required to investigate whether they are expressed in the same cell types and cellular location. Key to these investigations is the manufacture of a functioning NOD2 antibody. This is a problem that continues to hinder the scientific investigation of NOD2 in CD.

Chapter 7 Germline variation in MUC2 and MUC3A and association with IBD

Summary

Aims: To investigate if germline variation of the main mucin genes expressed in the gastrointestinal tract (*MUC2* and *MUC3A*) are associated with IBD susceptibility.

Methods: Tagging SNPs across the *MUC2* and *MUC3A* genes were selected and genotyped on the Taqman® platform in the Edinburgh IBD cohort of 446 CD, 452 UC and 428 controls.

Results: In *MUC2* the TG haplotype in block 1 (rs7942850T and rs11825977G) was significantly associated with controls compared with CD (haplotype frequency 0.469 in controls and 0.388 in CD, OR 0.72, p-value 0.0007). The SNP rs7942850C showed a statistically significant association with B2 disease (p=0.0027), with the TG haplotype including this SNP also demonstrating association with controls compared with B2 disease (p=0.0011). An analysis of association of *MUC2* and UC was negative. The TG haplotype in block 1 of *MUC2* was genotyped in the limited Dundee cohort of cases only, and demonstrated a haplotype frequency of 0.428. The *MUC3A* study was negative in both CD and UC.

Conclusion: Further genotyping of the rs7942850T and rs11825977G alleles in a larger cohort are required. Emerging evidence of the importance of MUC2 in inflammation and colon cancer makes mucins an important topic for future studies in IBD.

7.1 Introduction

The mucin genes are a family of related genes encoding apomucin proteins. Twenty-one mucins have been identified so far. Secretory mucins are clustered around chromosome 11p15.5[275] whereas membrane bound mucins are found over 3 regions: 7q22[276], 3q[277] and 1q21.[278] Following translation as apomucin proteins they are transferred to the Golgi apparatus where they undergo post translational modification in the form of O-glycosylation, involving many enzymes including those of the GALNT family. O-glycosylation is the attachment of glycans to serine and threonine amino acids on the apomucin. As the apomucins have serine and threonine rich amino acid sequence from DNA displaying variable number tandem repeats (VNTR), extensive O-glycosylation occurs (these areas are termed ‘mucin domains’). This gives mucins high molecular weights and the ability to retain water and become more resistant to proteolysis. In addition to these highly O-glycosylated sequences the mucin glycoprotein also has cysteine rich peptide sequences that are involved in the disulphide bond linkage between mucin subunits. As these areas are not O-glycosylated, they are susceptible to proteolytic attack.

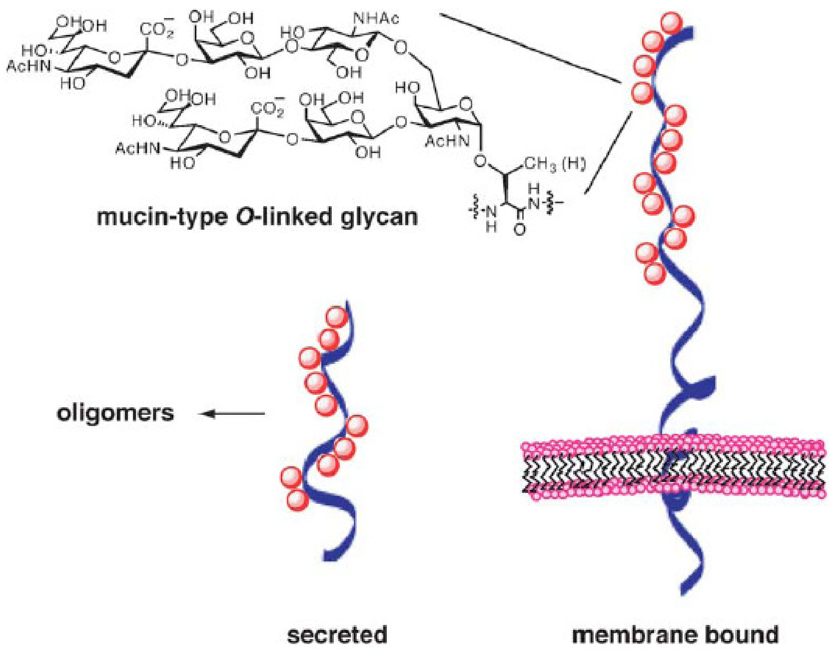


Figure 7-1 Structure of mucins, from Hang et al.[279]

Mucins occur in 2 forms, as illustrated in Figure 7-1:

1) Secretory, where the glycoprotein product is secreted into the lumen and forms part of the loosely bound mucus layer and

2) Membrane-bound: on the epithelial cell surface as part of the glycocalyx[280]

Mucin production is either due to constitutive baseline secretion or compound exocytosis, the latter being promoted by a variety of secretogogues.[281]

7.1.1 Mucins in the gastrointestinal tract

As discussed in Chapter 1, the mucosal defensive barrier is a vital part of gut barrier protection, as shown in Figure 7-2. Secreted mucins form an important part of the loosely adherent mucin layer protecting the gastrointestinal tract. Membrane bound mucins form part of the deeper, more adherent mucus layer. Mucins in the gastrointestinal tract are predominantly produced by goblet cells. MUC2 and MUC5 are the main secreted mucins expressed in the gastrointestinal tract, with MUC2 predominant in the small and large intestines.[210;282;283] Of the membrane associated mucins MUC1, MUC3A, MUC12, MUC13 and MUC17 appear to be expressed in the gastrointestinal tract[280;283], with MUC3A being predominant.[210] A study examining intestinal biopsies used real-time quantitative RT-PCR to look at mucin expression in the gut[280] demonstrated that the predominant mucins in the ileum were MUC2, MUC13 and MUC17, whereas in the colon it was MUC2, MUC12, MUC13 and MUC17. MUC10 and MUC19 were not covered by probesets on the DNA microarray. MUC3 was not mentioned in the paper for reasons that are not entirely clear.

MUC2 is a secreted glycoprotein which is strongly expressed by goblet cells in the bowel. The gene codes for a protein core of about 5100 amino acids.[282] The MUC3 glycoproteins are transmembrane mucins.[284] They consist of 2 genes: *MUC3A* and *MUC3B*, both on chromosome 7q22[285;286] which are expressed in the small intestine and colon.[285] MUC12[287], MUC13[288] and MUC17[289] are all transmembrane mucins encoded by genes on chromosome 7q22 which have expression in the GI tract. MUC19 is a secreted mucin with predominant expression in the submandibular gland[290]; it is not known to be expressed in the gastrointestinal tract.

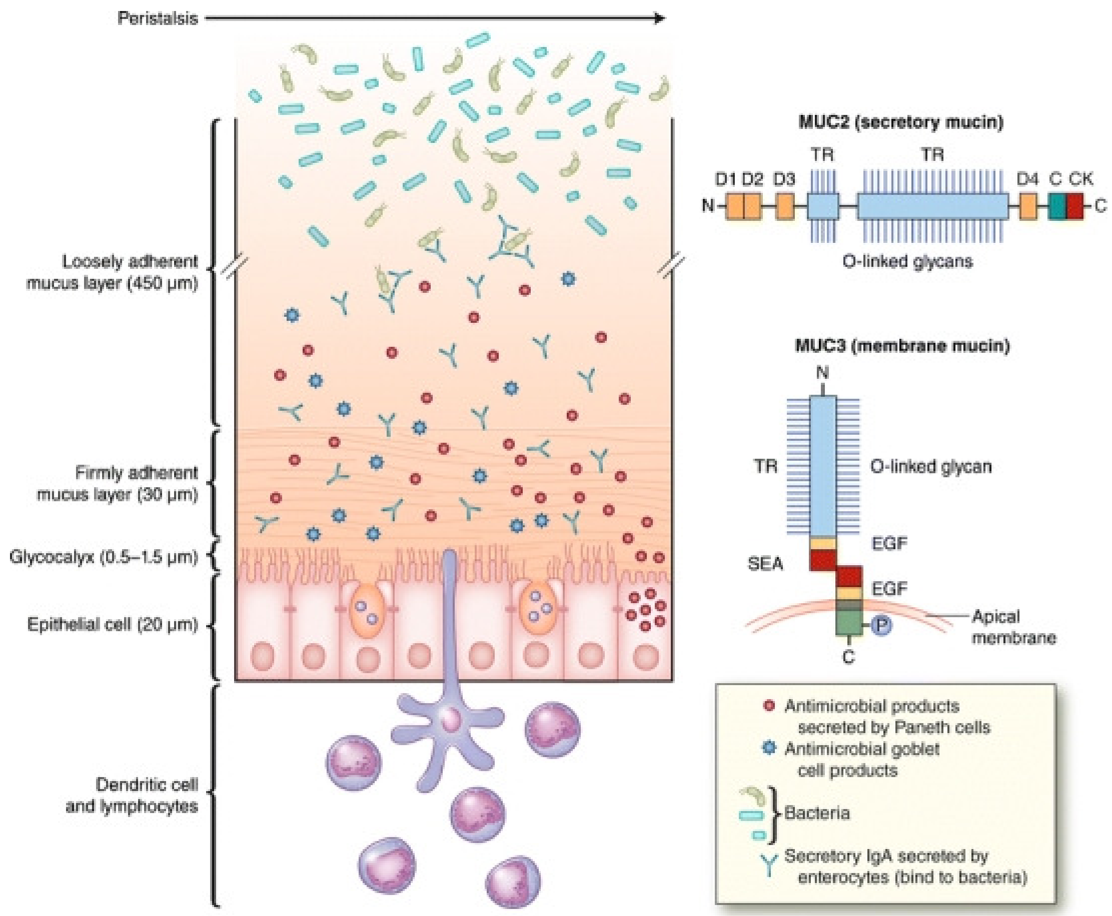


Figure 7-2 Gut mucus layers and domain structures of MUC2 and MUC3. TR=tandem repeat domain, D1-4=von Willebrand factor domain, CK=C-terminal cysteine knot domain, EGF=Epidermal growth factor-like domain, SEA=sperm protein, enterokinase and agrin domain. From Kim and Ho[291]

7.1.2 Mucins and association with cancer

Changes in mucin expression have long been known to be associated with colonic adenocarcinomas. Reduced MUC2 and MUC3 expression has been observed in human non-mucinous colon cancer specimens[292;293] and Muc2 $^{-/-}$ mice have a tendency to develop small intestine and large intestine adenomas, suggesting that it may be involved in the suppression of colorectal cancers.[294] MUC2 expression is correlated with less aggressive tumour behaviour.[295]

7.1.3 Mucins and IBD

In a dextran sodium sulfate (DSS) induced colitis mouse model, colonic *Muc2* mRNA expression is reduced in both acute and chronic colitis, whereas *Muc3* mRNA expression is increased in acute colitis but returns to basal levels in chronic colitis.[296] A similar rat model looking at small intestine and colonic mucin expression indicated that *Muc2* mRNA expression was reduced in the ileum but unchanged in the colon in acute DSS induced colitis, whereas *Muc2* mRNA expression was unchanged in the ileum and increased in the colon in acute colitis.[297] Muc2 $^{-/-}$ mice - and even Muc2 $^{+/-}$ to an extent - are highly susceptible to DSS induced colitis compared with Muc2 wild type mice.[298]

IBD, particularly UC[299;300], is characterised by a reduced layer of mucin in the areas of the bowel where there is inflammation, although whether this is cause or effect is not understood. Goblet cell depletion is recognised as being one of the hallmarks of UC.[301] There is evidence to suggest that mucin mRNA expression[280] especially *MUC2* mRNA[210] and MUC2 secretion[302] is down-regulated in the UC inflamed gut.

7.1.4 MUC2 gene

MUC2 is a gene on chromosome 11p15.5 spanning 30kbp. It consists of 49 exons including a large central exon containing tandem repeat sequences and encodes an apomucin of 5100 amino acids[282], the structure of which is shown in Figure 7-2.

7.1.5 MUC3A gene

MUC3A is located on chromosome 7q22 and consists of at least 12 exons. The gene encodes a transmembrane apoprotein containing a large extracellular domain, an epidermal growth factor-like transmembrane domain and a cytoplasmic tail[303], as illustrated in Figure 7-2. It contains a VNTR region that has been incompletely sequenced so far.

Early studies, before the advent of GWAS, indicated linkage between *MUC3* and IBD. A linkage study using microsatellite markers in 186 sibling pairs had demonstrated a lod (logarithm of odds) score of 3.08 for RFLP marker D7S669, located within 25Mbp of the *MUC3* gene locus.[304] This was further examined in Japanese and Caucasian UC and control populations by using a Southern blot analysis of *MUC3* alleles following restriction enzyme digestion.[286] This indicated that rare VNTR *MUC3* alleles were associated with IBD. This study was subsequently expanded to include CD patients and it was discovered that in fact

MUC3 consisted of 2 transcripts, *MUC3A* and *MUC3B*, with rare *MUC3A* exonic variants being identified with IBD more often than in controls but no *MUC3B* variants showing association.[303] However no individual alleles showed association with IBD overall or CD or UC separately. This study was limited by the technology available at the time, and the small size of the cohorts.

7.1.6 Hypothesis

It was hypothesized that germ line changes in mucin genes expressed in the gastrointestinal tract may be associated with increased IBD susceptibility. The *MUC2* and *MUC3A* genes were examined as they are the major intestinal mucins.

7.2 Methods

7.2.1 MUC2 genotyping

SNP data for *MUC2* was downloaded from www.hapmap.org, including the 5Kbp regions upstream and downstream of the gene. Tagging SNPs were selected across the gene (tagging for haplotype frequency >5%) as shown in Figure 7-3 and genotyped in the Edinburgh cohort of 446 CD, 452 UC and 428 controls on the Taqman® platform. This required 12 SNPs, as listed in Table 7-1.

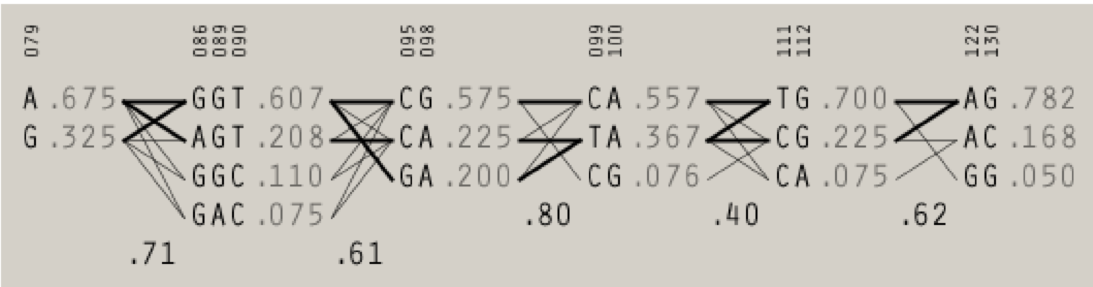


Figure 7-3 MUC2 Haplotypes >5%

SNP	Chr 11 position B36
rs7942850	1058900
rs11825977	1065920
rs11245936	1074362
rs10794288	1074811
rs7944723	1083710
rs6421972	1086494
rs10794293	1088939
rs11245954	1091078
rs11245962	1101164
rs6597976	1101474
rs7479605	1110324
rs7105198	1130041

Table 7-1 MUC2 SNP selection

7.2.2 MUC3A genotyping

SNP data for *MUC3A* was downloaded from www.hapmap.org, covering the 3 *MUC3A* transcripts and allowing for 15kbp each side. When the data was run through Haploview, blocks defined according to solid spine of LD and haplotypes examined above 5%, there were a large number of variants, as shown in Figure 7-4. In order to adequately tag each block, a total of 13 SNPs would have been needed.

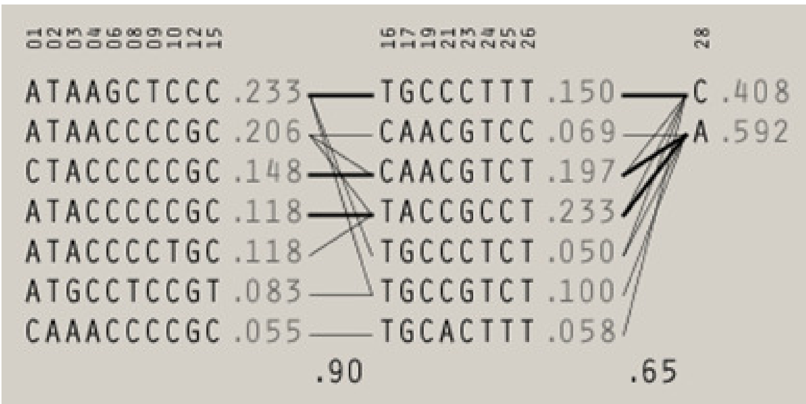


Figure 7-4 MUC3A Haplotypes >5% frequency

In order to reduce the number of SNPs required, only haplotypes above 10% were examined. This meant that blocks one and two could be joined, with the resultant blocks as shown in Figure 7-5. This meant that only 3 SNPs were required to tag the gene adequately.

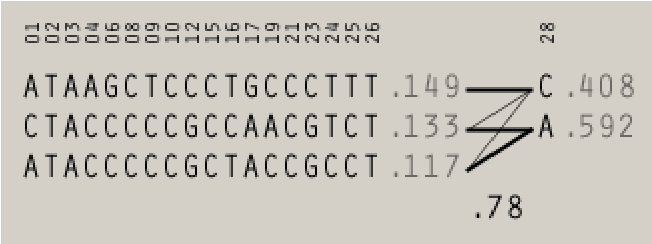


Figure 7-5 MUC3A Haplotypes >10% frequency

The 3 SNPs: rs4341099, rs11762787 and rs11765099, were genotyped on the Taqman® platform in the Edinburgh cohort of 446 CD, 452 UC and 428 controls.

7.3 Results

7.3.1 MUC2 genotyping

7.3.1.1 Quality control

All the SNPs were in Hardy-Weinberg equilibrium in controls.

7.3.1.2 MUC2 single SNP and haplotype analysis

When corrected for multiple testing across the 12 SNPs, a significant p-value was defined as <0.0042. The LD plots for CD and UC are shown in Figure 7-6. A single SNP analysis across *MUC2* showed no association with CD or UC, as shown in Table 7-2. Results of a haplotype analysis are shown in Table 7-3. The haplotypes were defined based on the haplotypes on which the SNPs had initially been chosen. The TG haplotype in block 1 was significantly associated with controls, with a p-value of 0.0007, and MAF of 0.469 in controls and 0.388 in CD. This conferred an OR of 0.72 (95% CI 0.59-0.87). No other haplotypes reached statistical significance after correction for multiple testing.

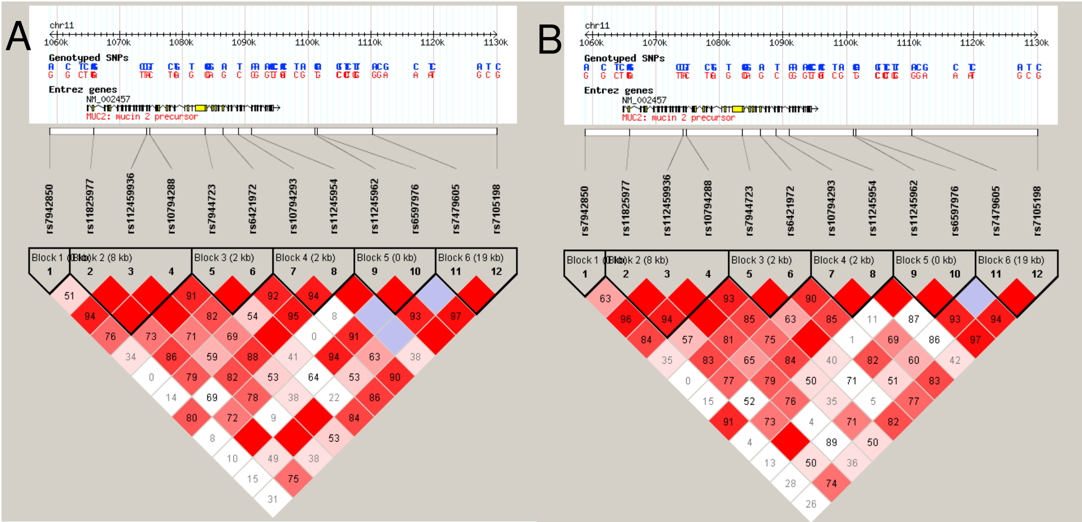


Figure 7-6 Haploview LD plots for CD (A) and UC (B)

	Minor allele	Control MAF	IBD MAF	p-value	CD MAF	p-value	UC MAF	p-value
rs7942850	C	0.378	0.409	0.148	0.443	0.0081	0.375	0.8875
rs11825977	A	0.176	0.211	0.0457	0.224	0.0171	0.198	0.2689
rs11245936	A	0.106	0.1	0.6199	0.097	0.5073	0.103	0.8418
rs10794288	C	0.212	0.19	0.1915	0.174	0.0489	0.206	0.7555
rs7944723	C	0.208	0.194	0.4168	0.185	0.2471	0.203	0.7939
rs6421972	T	0.409	0.423	0.5105	0.43	0.3952	0.416	0.7657
rs10794293	T	0.362	0.373	0.607	0.377	0.5217	0.368	0.7966
rs11245954	G	0.081	0.078	0.7719	0.073	0.5677	0.082	0.9502
rs11245962	C	0.336	0.294	0.0389	0.276	0.0105	0.312	0.2997
rs6597976	T	0.069	0.055	0.1649	0.049	0.0976	0.06	0.4646
rs7479605	C	0.088	0.095	0.5711	0.1	0.3886	0.089	0.9023
rs7105198	G	0.16	0.151	0.5767	0.138	0.2209	0.164	0.8284

Table 7-2 MUC2 Single SNP analysis

	Controls	IBD	p-value	CD	p-value	UC	p-value
Block 1							
TG	0.469	0.422	0.0246	0.388	7.00E-04	0.455	0.5982
CG	0.353	0.368	0.4701	0.39	0.1137	0.347	0.7305
TA	0.152	0.17	0.2262	0.171	0.2574	0.171	0.3341
CA	0.027	0.04	0.0846	0.051	0.012	0.027	0.6539
Block 2							
GTG	0.581	0.619	0.067	0.643	0.0091	0.594	0.554
GTC	0.206	0.191	0.3737	0.183	0.2234	0.2	0.7308
ACG	0.106	0.1	0.6196	0.097	0.5186	0.103	0.8287
GCG	0.105	0.087	0.1357	0.075	0.0279	0.1	0.6687
Block 3							
CCA	0.494	0.485	0.6861	0.488	0.7777	0.482	0.6835
TTA	0.342	0.357	0.4422	0.368	0.2721	0.347	0.8053
CCG	0.076	0.073	0.7845	0.07	0.6669	0.077	0.9512
TCA	0.065	0.063	0.8323	0.06	0.6114	0.067	0.8927
CTA	0.019	0.017	0.7392	0.012	0.3389	0.022	0.8553
Block 4							
TCTC	0.586	0.617	0.1252	0.632	0.0586	0.602	0.4749
CCTC	0.165	0.134	0.0376	0.13	0.0399	0.139	0.1314
CCTG	0.091	0.093	0.8336	0.086	0.7308	0.101	0.4874
TCCC	0.086	0.094	0.4887	0.099	0.291	0.089	0.7883
CTTG	0.068	0.055	0.189	0.051	0.1252	0.059	0.4507

Table 7-3 MUC2 Haplotype analysis

7.3.1.3 MUC2 CD sub phenotypic analysis

Results of a CD sub phenotypic analysis are shown in Table 7-4 and Table 7-5; no association was demonstrated when corrected for multiple testing. The SNP rs7942850C showed a statistically significant association with B2 disease (p=0.0027), with the TG haplotype including this SNP also demonstrating association with controls compared with B2 disease (p=0.0011). A haplotype analysis also showed no evidence of association (Table 7-6 and Table 7-7).

	Minor allele	Control MAF	L1 MAF	p-value	L2 MAF	p-value	L3 MAF	p-value
rs7942850	C	0.378	0.454	0.0303	0.451	0.0406	0.447	0.1336
rs11825977	A	0.176	0.203	0.331	0.228	0.0707	0.25	0.0436
rs11245936	A	0.106	0.056	0.015	0.095	0.6088	0.12	0.6384
rs10794288	C	0.212	0.136	0.0069	0.187	0.3836	0.203	0.811
rs7944723	C	0.208	0.205	0.9058	0.194	0.6221	0.15	0.1126
rs6421972	T	0.409	0.407	0.9571	0.451	0.2382	0.44	0.4936
rs10794293	T	0.362	0.376	0.6864	0.393	0.3773	0.336	0.568
rs11245954	G	0.081	0.057	0.2056	0.072	0.6466	0.079	0.9248
rs11245962	C	0.336	0.278	0.0904	0.286	0.1477	0.25	0.0542
rs6597976	T	0.069	0.038	0.0736	0.048	0.2429	0.082	0.5727
rs7479605	C	0.088	0.117	0.1701	0.089	0.9307	0.116	0.2888
rs7105198	G	0.16	0.142	0.4932	0.124	0.17	0.157	0.9327

Table 7-4 MUC2 CD sub phenotypic analysis – disease location

	Minor allele	Control MAF	B3 MAF	p-value	B2 MAF	p-value	B1 MAF	p-value
rs7942850	C	0.378	0.49	0.0275	0.527	0.0027	0.426	0.1079
rs11825977	A	0.176	0.179	0.9379	0.222	0.2438	0.233	0.0193
rs11245936	A	0.106	0.045	0.0398	0.045	0.0444	0.096	0.5541
rs10794288	C	0.212	0.116	0.0175	0.094	0.0043	0.188	0.3143
rs7944723	C	0.208	0.202	0.8787	0.191	0.6782	0.183	0.2999
rs6421972	T	0.409	0.387	0.6634	0.443	0.4971	0.436	0.3598
rs10794293	T	0.362	0.296	0.1804	0.441	0.1194	0.378	0.585
rs11245954	G	0.081	0.045	0.1755	0.056	0.3552	0.075	0.7272
rs11245962	C	0.336	0.192	0.0032	0.245	0.0657	0.295	0.1569
rs6597976	T	0.069	0.037	0.2106	0.046	0.3798	0.051	0.2282
rs7479605	C	0.088	0.111	0.4255	0.069	0.5186	0.11	0.2157
rs7105198	G	0.16	0.074	0.0193	0.12	0.2902	0.151	0.6978

Table 7-5 CD sub phenotypic analysis - disease behaviour

	Controls	L1	p-value	L2	p-value	L3	p-value
Block 1							
TG	0.469	0.393	0.0295	0.379	0.0112	0.369	0.0289
CG	0.353	0.405	0.1305	0.395	0.2181	0.387	0.4532
TA	0.152	0.156	0.8706	0.172	0.4248	0.188	0.2786
CA	0.027	0.047	0.0875	0.053	0.0328	0.056	0.0502
Block 2							
GTG	0.581	0.657	0.0229	0.619	0.2542	0.65	0.1173
GTC	0.206	0.205	0.9251	0.195	0.6543	0.144	0.0797
GCG	0.106	0.08	0.2093	0.091	0.4692	0.08	0.3486
ACG	0.105	0.057	0.0165	0.094	0.5988	0.119	0.6413
Block 3							
CCA	0.494	0.516	0.5379	0.475	0.5824	0.485	0.8408
TTA	0.342	0.361	0.5669	0.383	0.2284	0.341	0.9919
CCG	0.076	0.056	0.2714	0.069	0.7101	0.076	0.9945
TCA	0.065	0.052	0.4119	0.059	0.7161	0.093	0.2475
CTA	0.019	0.014	0.6452	0.011	0.4844	0.003	0.1854
Block 4							
TCTC	0.586	0.612	0.4807	0.633	0.1917	0.641	0.2185
CCTC	0.165	0.135	0.2249	0.153	0.6176	0.083	0.0122
CCTG	0.091	0.1	0.6593	0.073	0.3811	0.077	0.6003
TCCC	0.086	0.112	0.1627	0.088	0.7918	0.115	0.2296
CTTG	0.068	0.039	0.0848	0.051	0.301	0.076	0.7451

Table 7-6 MUC2 Crohn's sub phenotypic haplotype analysis – disease location

	Controls	B3	p-value	B2	p-value	B1	p-value
Block 1							
TG	0.469	0.384	0.094	0.303	0.0011	0.4	0.0177
CG	0.353	0.438	0.0879	0.476	0.0137	0.37	0.5286
TA	0.152	0.134	0.6078	0.17	0.6513	0.175	0.2456
CA	0.027	0.045	0.2275	0.052	0.102	0.055	0.0162
Block 2							
GTG	0.581	0.682	0.036	0.713	0.0072	0.629	0.0952
GTC	0.206	0.202	0.8766	0.191	0.6766	0.182	0.2914
GCG	0.106	0.071	0.2401	0.051	0.0654	0.091	0.5594
ACG	0.105	0.045	0.0404	0.045	0.0457	0.095	0.3988
Block 3							
CCA	0.494	0.557	0.2112	0.48	0.7899	0.484	0.7249
TTA	0.342	0.291	0.2846	0.419	0.1102	0.37	0.3197
CCG	0.076	0.043	0.2095	0.056	0.4402	0.073	0.8414
TCA	0.065	0.096	0.2429	0.024	0.0865	0.063	0.8447
CTA	0.019	0.012	0.6366	0.021	0.8651	0.008	0.1704
Block 4							
TCTC	0.586	0.702	0.0202	0.687	0.049	0.604	0.5236
CCTC	0.165	0.114	0.1628	0.125	0.2897	0.135	0.1608
CCTG	0.091	0.037	0.057	0.069	0.4556	0.098	0.6826
TCCC	0.086	0.107	0.3692	0.063	0.5152	0.109	0.1812
CTTG	0.068	0.037	0.1965	0.049	0.4448	0.051	0.2288

Table 7-7 MUC2 Crohn's sub phenotypic haplotypic analysis - disease behaviour

7.3.1.4 UC sub phenotypic analysis

Results of a UC sub phenotypic analysis are shown in Table 7-8 and Table 7-9, and were negative when corrected for multiple testing.

SNP	Minor allele	Control MAF	E3 MAF	p-value	E2 MAF	p-value	E1 MAF	p-value
rs7942850	C	0.378	0.36	0.5367	0.396	0.5937	0.355	0.614
rs11825977	A	0.176	0.178	0.9269	0.194	0.4803	0.262	0.0231
rs11245936	A	0.106	0.104	0.8984	0.092	0.4659	0.117	0.7156
rs10794288	C	0.212	0.213	0.9593	0.202	0.7081	0.192	0.6103
rs7944723	C	0.208	0.199	0.7174	0.191	0.5294	0.219	0.7795
rs6421972	T	0.409	0.378	0.3117	0.425	0.6197	0.492	0.0783
rs10794293	T	0.362	0.329	0.2655	0.383	0.5153	0.444	0.0757
rs11245954	G	0.081	0.092	0.5161	0.069	0.4894	0.077	0.8762
rs11245962	C	0.336	0.317	0.5245	0.322	0.6509	0.254	0.0728
rs6597976	T	0.069	0.058	0.4706	0.07	0.9528	0.04	0.2332
rs7479605	C	0.088	0.08	0.6731	0.086	0.9307	0.119	0.2578
rs7105198	G	0.16	0.173	0.567	0.169	0.7067	0.133	0.4379

Table 7-8 MUC2 UC sub phenotypic analysis

	Controls	E3	p-value	E2	p-value	E1	p-value
Block 1							
TG	0.469	0.482	0.6319	0.436	0.3449	0.427	0.3913
CG	0.353	0.339	0.5691	0.368	0.695	0.314	0.3632
TA	0.152	0.157	0.8627	0.169	0.5159	0.22	0.0559
CA	0.027	0.022	0.8405	0.026	0.7683	0.039	0.3283
Block 2							
GTG	0.581	0.594	0.6634	0.607	0.3757	0.588	0.8431
GTC	0.206	0.192	0.5674	0.192	0.5455	0.219	0.7671
ACG	0.105	0.104	0.9035	0.092	0.4746	0.115	0.7465
GCG	0.106	0.103	0.9004	0.109	0.9392	0.077	0.295
Block 3							
CCA	0.494	0.509	0.6078	0.485	0.8061	0.414	0.092
TTA	0.342	0.304	0.2048	0.364	0.4799	0.426	0.0604
CCG	0.076	0.084	0.6498	0.066	0.5193	0.077	0.9935
TCA	0.065	0.07	0.7761	0.062	0.8081	0.066	0.9904
CTA	0.019	0.025	0.564	0.021	0.8724	0.017	0.9261
Block 4							
TCTC	0.586	0.608	0.4541	0.595	0.7745	0.633	0.3314
CCTC	0.165	0.133	0.1362	0.147	0.442	0.114	0.1404
CCTG	0.091	0.114	0.1986	0.092	0.9318	0.092	0.9556
TCCC	0.086	0.082	0.8465	0.082	0.9361	0.108	0.334
CTTG	0.068	0.057	0.4859	0.07	0.8919	0.034	0.1335

Table 7-9 MUC2 UC sub phenotypic haplotype analysis

7.3.1.5 Replication

As the TG haplotype in block 1 of the *MUC2* genotyping demonstrated a positive p-value in CD when compared with controls, it merited an attempt to confirm the findings in a separate cohort of patients. Unfortunately, we were only able to attempt to replicate the CD MAF in the small Dundee cohort of 299 CD and 213 UC samples. The comparisons of haplotype frequency are shown in Table 7-10. When the Dundee samples were compared directly with the Edinburgh controls, there was no significant difference between the 2 cohorts.

	Dundee CD	Edinburgh CD	Edinburgh controls
TG	0.428	0.388	0.469
CG	0.371	0.39	0.353
TA	0.184	0.171	0.152
CA	0.017	0.051	0.027

Table 7-10 Haplotype frequency in different cohorts

7.3.2 MUC3A genotype

The 3 SNPs were in Hardy Weinberg equilibrium in controls. A significant p-value was defined as <0.016 when corrected for multiple testing. Results of a *MUC3A* single SNP and haplotype analysis are shown in Table 7-11 and Table 7-12, and were negative. The study had approximately 70% power to detect OR>1.4, and >99% power to detect OR of 2.6 (OR given in the previous MUC3A study[286]) Given the negative results and lack of trends to significance, a sub phenotypic analysis was not completed.

	Minor allele	Control MAF	IBD MAF	p-value	CD MAF	p-value	UC MAF	p-value
rs11762787	C	0.26	0.257	0.8585	0.264	0.8331	0.25	0.6323
rs4341099	C	0.259	0.281	0.2424	0.298	0.0828	0.268	0.6889
rs11765099	C	0.463	0.455	0.7332	0.465	0.9289	0.448	0.5253

Table 7-11 MUC3A Single SNP analysis

	Controls	IBD	p-value	CD	p-value	UC	p-value
Block 1							
GT	0.479	0.461	0.4005	0.437	0.0883	0.482	0.9162
GC	0.26	0.282	0.2388	0.299	0.0784	0.268	0.6932
CT	0.262	0.257	0.8076	0.265	0.8925	0.251	0.6036
Block 2							
A	0.537	0.544	0.737	0.535	0.9299	0.552	0.5312
C	0.463	0.456	0.737	0.465	0.9299	0.448	0.5312

Table 7-12 MUC3A Haplotype analysis

7.4 Discussion

A definite link between the *MUC2/MUC3A* genes and IBD susceptibility has not been demonstrated by the data presented in this chapter.

7.4.1 Mucins and genetics

With the SNP rs7942850 and its two-marker associated haplotype showing association in the Edinburgh CD cohort, it is certainly possible that *MUC2* is indeed an IBD susceptibility gene. The replication study did not have a control group that was genotyped to make it a formal replication cohort. With the bigger Dundee cohort now available, including controls, there is a plan to complete genotyping on the full Dundee cohort. However, if there is lack of an association in this cohort, it will still not be possible to refute the association, as it would be underpowered for the OR in the discovery cohort. For an OR of 1.39 and an allelic frequency of 0.40, at least 650 case-control pairs would be required to rule-out an association with 80% power.

The *MUC3A* case-control study was negative. Limitations of the study include the small cohort size, the tagging SNPs being chosen on the basis of a haplotype frequency of >10% (rather than 5%) as well as the fact that there is incomplete data on the correct *MUC3A* sequence (due to the VNTR region being incompletely sequenced). The initial interest in *MUC3A* was due to the finding that a lod score of 3.08 for the D7S669 RFLP marker (‘IBD11’). However, in the era of more detailed chromosomal mapping, this area is actually more than 20Mbp upstream of the *MUC3A* gene. A more recent large UC GWAS has re-awakened interest in the area with a SNP (rs7809799) 2Mbp upstream of *MUC3A*, in between the *SMURF1* and *KPNA7* genes, conferring an OR of 1.56 for UC susceptibility.[77] This area, rather than *MUC3A*, appears more likely to be responsible for the signal at this locus.

The traditional candidate gene approach to gene discovery has been superseded by GWAS which reduce the risk of false positive gene susceptibility studies but may increase the rate of false negatives due to population pooling. It would be useful to download the relevant WTCCC GWAS data to analyse for gene-wide significance for any SNPs within these genes. Certainly there is no genome-wide significance between *MUC2* or *MUC3A* and IBD susceptibility on any recent GWAS.

Even when studies are negative overall for CD, UC and IBD, sub phenotypic studies are still worthwhile as a gene may be solely associated with a subtype of disease (for example, L1 disease). This is why the relevant analyses are detailed in this chapter.

7.4.2 Cause or effect?

Another consideration is whether the reduced mucin layer and goblet cell reduction in UC is the cause or the effect of the disease. The mucin layer could be reduced or functioning abnormally contributing to bacterial damage to the epithelial layer. Or it could be that the inflammation and damage to goblet cells is causing reduced mucin production.

Even given the former possibility, there are a number of ways that reduced mucin/abnormal mucin function could occur, aside from germline variation in the apomucin gene itself, as detailed in section 7.4.3.

7.4.3 Alternative causes for mucin abnormalities

Epigenetic changes in the mucin genes could change their expression. *MUC2*[305] and *MUC3A*[306] expression in colon cancer cell lines has been shown to be highly controlled by methylation and histone deacetylation. A recent study with the LS174T cell line has indicated that *MUC2* expression is regulated in part by short-chain fatty acids (SCFA), with butyrate and propionate inducing an increase in *MUC2* mRNA levels via histone acetylation/methylation and butyrate-responsive regions in the promoter area of the gene.[307] This is especially interesting as SCFA are the fermentation products of bacteria, so changes in intraluminal bacteria, by changing the SCFA composition of the luminal contents, may affect mucin gene expression. Of course, cancer cell lines can function very differently to their non-cancerous counterparts, but it still shows the potential importance of epigenetics in mucin regulation.

Bacteria affect mucin expression in other ways. *MUC2* has NF-kappaβ binding sites in its promoter region[308] and it has been demonstrated that TNFα[309;310]and LPS[308] activate *MUC2* transcription via NF-kappaβ in colonic cell lines.

A further mechanism for changes in mucin production could be in the post-translational processing of mucins, in particular O-glycosylation. The GALNT family of enzymes catalyse the first step in O-glycosylation in the Golgi apparatus - in Chapter 5, germline variation in *GALNT2* and the possible linkage with IBD susceptibility has been discussed. However, there are numerous other genes encoding enzymes and other regulators of O-glycosylation; germline variation or changes in expression in any one of them could affect the quality or quantity of mucin produced by the cell.

Mucin sialylation and sulfonation (addition of sialyl and sulfonyl groups respectively) are further post translational modifications that help to protect mucins from degradation, and changes or abnormalities in these processes could render mucins to be less effective in their role as protection against the luminal contents. In UC, there is evidence for reduced sulfation[311], although evidence for reduced sialylation is less clear-cut.

Mucolytics (eg bacterial glycosidases) produced by luminal bacteria themselves could also cause a reduction or change in the mucin layer, and increased ability to damage the underlying structures (reviewed by Deplancke and Gaskins[281]). A recent study had used 16S RNA technology to demonstrate that there are increased mucolytic bacteria in the mucin of IBD patients, but especially in CD patients.[312]

7.4.4 Conclusion

Despite the lack of definite evidence linking germline variation of IBD with *MUC2* and *MUC3A*, mucins in IBD still appears to be a highly relevant topic for further research, not only for the reasons given above, but also the fact that mucins could be a common link between IBD and colon cancer.

It would be interesting to study the epigenetics of *MUC2* and *MUC3A* in IBD patients and compare them with controls. Further bacterial studies examining in more detail the effects of specific bacteria – for example, adherent-invasive *E. Coli* – on mucin expression and production, would also be useful.

Much of the data on mucins and O-glycosylation in IBD is more than 10 years old; with increasing recognition not only of the part that luminal bacteria have to play in the pathogenesis of IBD but also of the role of the mucosal barrier in the defence against these bacteria, this topic is likely to be extensively studied in the future

Chapter 8 Germline variation in MUC19 and LRRK2 and association with IBD

Summary

Aims: To investigate whether either *MUC19* or *LRRK2* was responsible for the CD susceptibility signal at the rs11175593 locus.

Methods: *MUC19* tagging SNPs were genotyped on the Taqman® platform in the Edinburgh cohort of 437CD, 451 UC and 428 controls. *LRRK2* tagging SNPs were genotyped on the Sequenom® platform in the Dundee cohort of 366 CD, 261 UC and 539 controls.

Results: Genotyping across the *MUC19* gene was negative in IBD overall, and CD and UC separately, including appropriate sub phenotypic analyses of association. There were no statistically significant associations of *LRRK2* and IBD when corrected for multiple testing but there were multiple SNPs and haplotypes that demonstrated a trend to statistical significance.

Conclusion: *LRRK2* is most likely to represent the CD susceptibility gene at the rs11175593 locus, but larger adequately powered replication studies are required.

8.1 Introduction

As already described, GWAS in IBD have uncovered a large number of potential susceptibility genes. Combining cohorts of patients provides increased power to detect susceptibility genes conferring lower odds ratios (ORs), but potentially loses any population specific loci. The meta-analysis combining British, North American, Belgian-French CD GWAS[117] and the subsequent larger CD GWAS meta-analysis[74], has led to the discovery of several novel SNPs linked with CD susceptibility conferring ORs of 1.04 – 1.5. Many of these novel loci lie *near* rather than *within* genes of interest. This is the case for the strongest novel locus identified in the Barrett meta-analysis[117] which was tagged by the rs11175593T variant and estimated to confer an OR of 1.54 for CD susceptibility. This variant lies within the chromosome 12q12 region and is within 40Kbp (kilo base pairs) of the leucine-rich repeat kinase 2 (*LRRK2*) gene and 360Kbp of the *MUC19* gene (Figure 8-1).

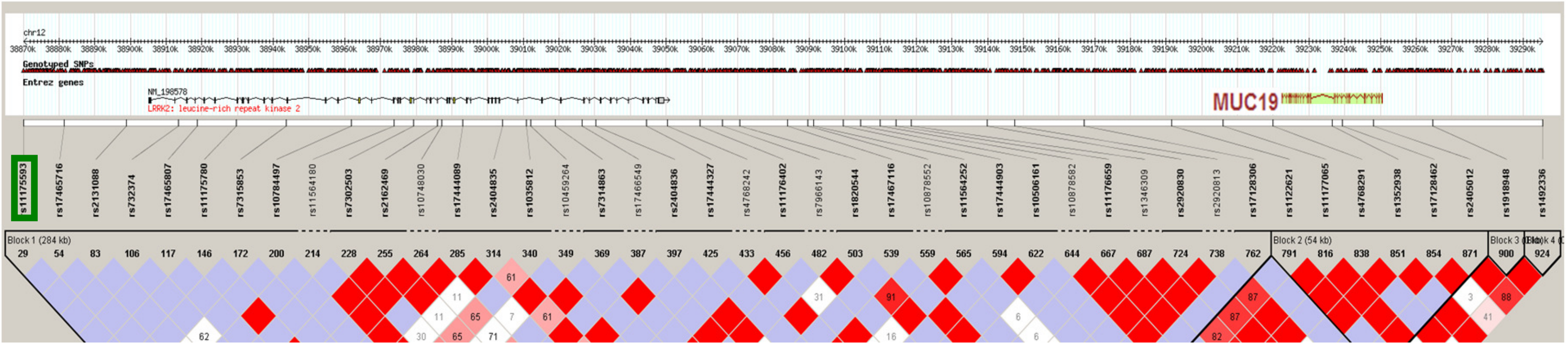


Figure 8-1 Haploview diagram of rs11175593 (green box) relative to LRRK2 & MUC19 (selected SNPs only shown)

8.1.1 MUC19

The family of *MUC* genes encode apoproteins that undergo post translational modification to become glycoproteins forming an important part of the barrier protection of epithelial cell surfaces.[313] MUC19 is a secreted mucin; it is the largest mucin gene identified to date[290] spanning more than 180kbp, with a deduced peptide sequence of >7000 amino acids. Although mRNA *MUC19* expression has been characterised in the mucous cells of human submandibular gland[290] and other species have demonstrated MUC19 glycoprotein in saliva, the MUC19 glycoprotein has not been isolated from human saliva.[314] *MUC19* mRNA transcripts and MUC19 expression on IHC have also been noted in the human cornea, conjunctiva and lacrimal gland[315] and mRNA transcripts found in the human middle ear.[316] There are no published data on *MUC19* expression in the gastrointestinal tract or in immune cells.

8.1.2 LRRK2

The *LRRK2* gene encodes a protein of almost 2500 amino acids[317] which is ubiquitously expressed.[318] It was first cloned in studies of familial Parkinson's disease (PD) where the signal at the PD susceptibility locus *PARK8* was linked with mutations in this gene.[317;318] As these cases of familial PD are inherited in an autosomal dominant pattern, it appears that the mutations cause a gain in function. LRRK2 is a cytoplasmic protein kinase that also associates with the mitochondria.[319] The gene contains a Ras of Complex Proteins (ROC) domain which acts as a GTPase to regulate its protein kinase activity. Most of the research in LRRK2 has been in the context of PD, where the mutant form of the protein increased apoptotic cell death and reduced neuronal cell viability in cell lines.[319] A neuroblastoma cell line transfected with *LRRK2* cDNA containing the most common PARK8 mutation, G2019S, resulted in significant reduction in neurite length compared with cells transfected with wild-type *LRRK2*, and these G2019S-transfected cells showed markedly increased autophagic vacuoles.[320] These effects were reversed when LC3 and Atg7, two important components of autophagy, were knocked down by RNA interference in the G2019S *LRRK2* transfected cell line.[320] A further study suggested that LRRK2 is a negative regulator of autophagy activity, with LRRK2 knockdown increasing autophagic activity and impaired autophagic balance with a mutant LRRK2 form.[321]

LRRK2 has been detected in immune cells and expression is upregulated by IFNγ[322], a cytokine implicated in CD pathogenesis. In CD intestinal biopsy specimens, *LRRK2* mRNA expression is upregulated in inflamed tissues compared with uninflamed.[322] The same study also showed that LRRK2 activates NF-kappaB pathways, and that LRRK2 knock-down reduces killing of intracellular bacteria.[322]

8.1.3 Study aim

The aim of this study was to investigate whether either *MUC19* or *LRRK2* was responsible for the CD susceptibility signal at the rs11175593 locus.

8.2 Methods

8.2.1 MUC19

MUC19 SNPs were chosen using solid spine of linkage disequilibrium (LD) to tag haplotypic variation of the *MUC19* gene including the extended 5' and 3' regions (haplotype frequency >5%), and required three SNPs. These SNPs were genotyped on the Taqman® genotyping platform in the Edinburgh cohort of 437CD, 451 UC and 428 controls. Given that the case-control OR of association of rs11175593 with CD was 1.54, the study had more than 85% power to detect an association with CD for each of the 3 tagging variants.

8.2.2 LRRK2 SNP selection and genotyping

LRRK2 SNPs were chosen using solid spine of LD to tag haplotypic variation of the *LRRK2* gene, including the extended 5' and 3' regions, and genotyped on the Sequenom® platform at University of California, San Francisco in collaboration with Genentech, Inc. The haplotypic structure of the gene is shown in Figure 8-2. Due to the large number of combinations of SNP variants in the gene, the haplotype blocks were joined to tag for >5% variation. However this resulted in the SNP selection encompassing less variation of the gene. Thus an increased number of haplotype blocks, and therefore SNPs, were chosen (Figure 8-3, table 1-1). These 36 SNPs were genotyped in the Dundee cohort of 366 CD, 261 UC and 539 controls.

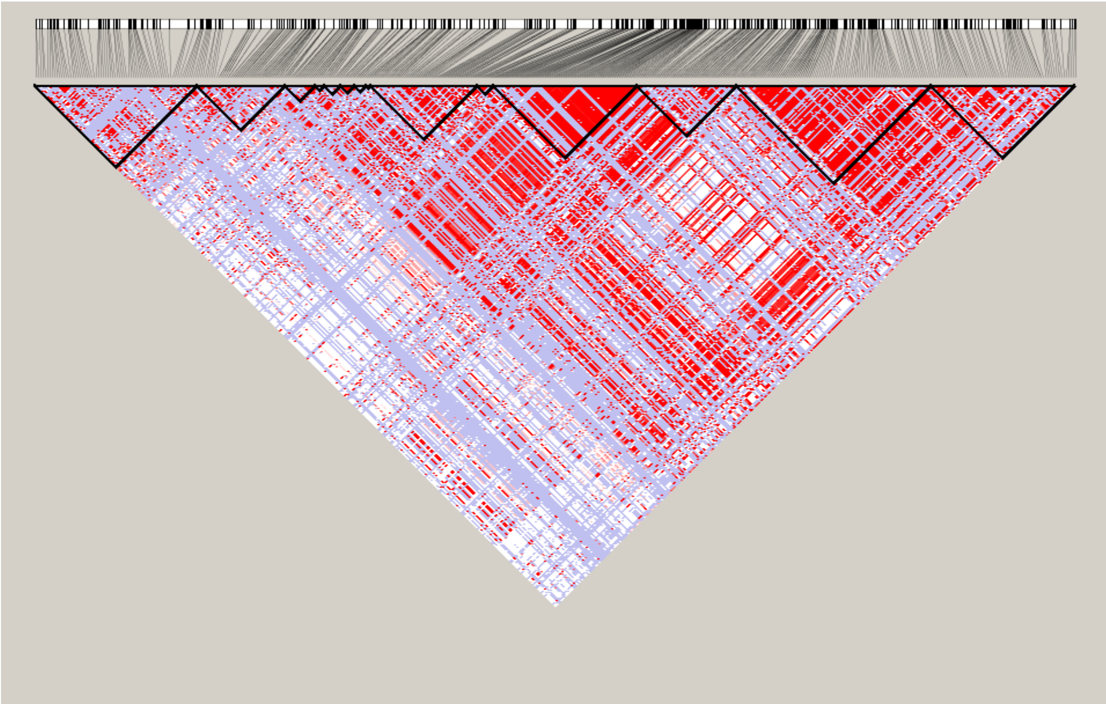


Figure 8-2 Haplotypic structure of LRRK2, with the haplotype blocks marked

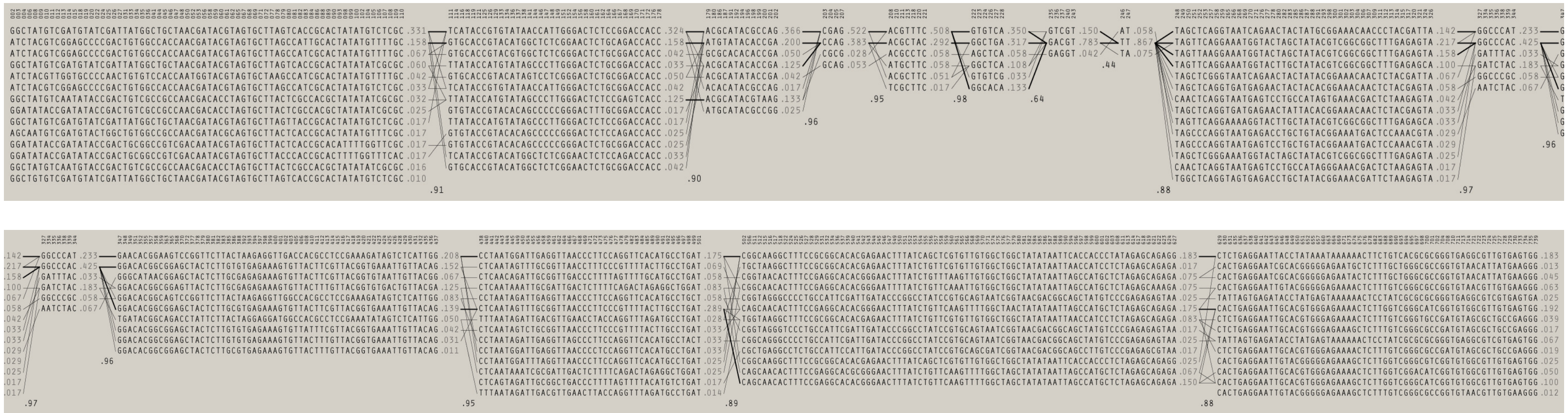


Figure 8-3 LRRK2 haplotypes on which SNP selection was made

SNP	Chr 12 Position Build 36	MAF in CEU HapMap population
rs7973254	38894033	0.35
rs12230685	38909057	0.175
rs10878246	38918366	0.183
rs6581622	38920425	0.3
rs10878258	38927959	0.292
rs7955902	38931524	0.383
rs2723264	38938787	0.25
rs17465912	38942662	0.133
rs10784462	38944038	0.442
rs11175784	38945801	0.492
rs10784470	38949863	0.383
rs7308193	38951494	0.317
rs10506150	38953027	0.15
rs10506151	38957265	0.158
rs7309197	38959527	0.458
rs17491061	38980285	0.133
rs10784499	38984757	0.283
rs10878343	38986851	0.233
rs11175958	38987655	0.233
rs11564128	38992102	0.125
rs2896975	38992642	0.325
rs10467144	38993468	0.383
rs11175985	38993545	0.225
rs4768230	39001569	0.392
rs10878371	39002527	0.383
rs11176030	39004720	0.392
rs2404832	39008719	0.283
rs17444124	39010761	0.475
rs6581667	39017773	0.433
rs10506155	39022206	0.317
rs2162471	39024099	0.217
rs10459265	39024570	0.433
rs4767973	39042515	0.325
rs11835105	39044312	0.2
rs17466605	39046247	0.333
rs17520676	39048180	0.206

Table 8-1 SNP selection for LRRK2

8.3 MUC19 Results

All three SNPs were in Hardy-Weinberg equilibrium in controls, and had >90% genotyping rates. After Bonferroni correction for multiple testing for the 3 SNPs at a p-value of 0.05, a significant p-value was <0.017. Single marker and haplotype susceptibility analyses did not show an association with any of the MUC19 haplotype-tagging variants in IBD overall, nor in CD or UC separately (Table 8-2). A detailed genotype-phenotype analysis for both CD and UC according to the Montreal classification found no associations in CD (Table 8-3: only L1, L2 and B3 comparisons with controls shown) and UC (Table 8-4: only E3 comparison with controls shown).

MUC 19 SNP tags	Minor allele	Control MAF	CD MAF	p-value	UC MAF	p-value
rs4768291	C	0.156	0.143	0.487	0.155	0.968
rs1352938	A	0.320	0.328	0.736	0.345	0.296
rs17128462	C	0.215	0.195	0.322	0.185	0.129

Table 8-2 Allelic frequencies and p-values for MUC19

	Minor allele	Controls	L1	p-value	L2	p-value	B3	p-value
rs4768291	C	0.156	0.159	0.880	0.137	0.353	0.143	0.711
rs1352938	A	0.320	0.307	0.625	0.320	0.990	0.347	0.560
rs17128462	C	0.215	0.197	0.434	0.194	0.373	0.172	0.265

Table 8-3 Allelic freq CD sub phenotypic analysis

	Minor allele	Controls	E3	p-value
rs4768291	C	0.156	0.185	0.202
rs1352938	A	0.320	0.325	0.869
rs17128462	C	0.215	0.182	0.191

Table 8-4 Allelic freq UC sub phenotypic analysis

Further detailed analysis of available HapMap data showed a complete lack of LD between rs11175593 and the three *MUC19* tagging variants (r^2=0).

8.4 LRRK2 results

8.4.1 Quality control

Two SNPs were discounted: rs11175958 (successful genotyping of < 90%) and rs10506151 (due to lack of Hardy-Weinberg equilibrium in the controls). Nineteen controls and three IBD DNAs were excluded due to unsuccessful genotyping (<90%).

8.4.2 Single SNP analysis

Single SNP marker analysis for association with IBD, CD or UC is shown in Table 8-5. When corrected for multiple testing, the level of significance was $p<0.0015$, which was not attained by any of the SNPs in IBD, CD or UC compared with controls. However, both rs10878246 and rs7955902 had borderline significance in IBD and UC (rs10878246: IBD $p=0.0175$, UC $p=0.0018$; rs7955902: IBD $p=0.0061$, UC $p=0.0164$). In CD patients none of the SNPs approached statistical significance.

Results of a sub phenotypic analysis for the single SNPs in CD are shown in Table 8-6 and Table 8-7. The SNP rs7955902 had a borderline significant association with L1 disease location ($p =0.004$). However, 13 of the 34 SNPs were associated with B3 internal penetrating disease compared with controls at a $p <0.05$ (Table 8-7), although none of the associations reached statistical significance overall.

Results of a sub phenotypic analysis for the single SNPs in UC are shown in Table 8-8. Five SNPs across the gene were associated with E3 disease (pancolitis) compared with controls ($p<0.05$). However, none reached statistical significance overall when corrected for multiple testing.

SNP	Minor allele	Control MAF	IBD MAF	p-value	CD MAF	p-value	UC MAF	p-value
rs7973254	C	0.315	0.316	0.9756	0.31	0.805	0.323	0.7646
rs12230685	T	0.119	0.123	0.7343	0.129	0.5142	0.12	0.958
rs10878246	G	0.185	0.225	0.0175	0.203	0.3329	0.253	0.0018
rs6581622	C	0.265	0.275	0.6069	0.282	0.449	0.268	0.899
rs10878258	G	0.264	0.264	0.969	0.275	0.6284	0.251	0.5696
rs7955902	A	0.479	0.421	0.0061	0.432	0.0599	0.413	0.0164
rs2723264	T	0.173	0.177	0.8453	0.191	0.3478	0.17	0.8608
rs17465912	T	0.151	0.171	0.2045	0.159	0.6468	0.191	0.0459
rs10784462	C	0.481	0.48	0.9711	0.486	0.8211	0.458	0.3858
rs11175784	C	0.509	0.507	0.953	0.503	0.6368	0.467	0.3652
rs10784470	T	0.357	0.334	0.2276	0.338	0.4014	0.335	0.3922
rs7308193	G	0.314	0.308	0.745	0.323	0.7012	0.295	0.4484
rs10506150	T	0.122	0.112	0.4772	0.109	0.3962	0.122	0.9988
rs7309197	T	0.49	0.493	0.8866	0.501	0.6498	0.469	0.4281
rs17491061	C	0.157	0.168	0.4555	0.155	0.9366	0.191	0.0939
rs10784499	T	0.337	0.363	0.1954	0.339	0.926	0.376	0.1255
rs10878343	T	0.204	0.206	0.9234	0.223	0.333	0.201	0.8805
rs11564128	T	0.147	0.17	0.1375	0.158	0.5423	0.191	0.0274
rs2896975	G	0.33	0.326	0.8441	0.346	0.4654	0.311	0.4572
rs10467144	A	0.432	0.436	0.8598	0.449	0.4939	0.422	0.6971
rs11175985	T	0.147	0.137	0.4708	0.126	0.217	0.154	0.6984
rs4768230	A	0.331	0.311	0.2862	0.313	0.4198	0.313	0.4582
rs10878371	T	0.426	0.438	0.5586	0.451	0.2976	0.424	0.9485
rs11176030	T	0.334	0.312	0.2465	0.315	0.3912	0.313	0.3998
rs2404832	G	0.338	0.363	0.2128	0.341	0.9116	0.375	0.1557
rs17444124	C	0.479	0.482	0.9213	0.473	0.7736	0.504	0.3659
rs6581667	G	0.47	0.456	0.4991	0.462	0.7159	0.444	0.3264
rs10506155	A	0.333	0.322	0.5985	0.326	0.784	0.324	0.741
rs2162471	C	0.178	0.182	0.8049	0.197	0.3075	0.177	0.9683
rs10459265	G	0.472	0.456	0.4418	0.462	0.6574	0.444	0.2925
rs4767973	A	0.311	0.309	0.9227	0.326	0.5261	0.29	0.3795
rs11835105	G	0.803	0.209	0.4811	0.199	0.9231	0.222	0.2556
rs17466605	A	0.313	0.311	0.9046	0.327	0.5425	0.292	0.3807
rs17520676	-	0	0		0		0	

Table 8-5 LRRK2 single SNP analysis for IBD, CD and UC vs controls, p-values<0.05 highlighted

SNP	Minor allele	Control MAF	L1 MAF	p-value	L2 MAF	p-value	L3 MAF	p-value
rs7973254	C	0.315	0.368	0.170	0.288	0.416	0.297	0.559
rs12230685	T	0.119	0.153	0.197	0.112	0.754	0.129	0.626
rs10878246	G	0.185	0.222	0.251	0.194	0.737	0.192	0.775
rs6581622	C	0.265	0.318	0.145	0.277	0.715	0.269	0.895
rs10878258	G	0.264	0.312	0.184	0.269	0.893	0.262	0.942
rs7955902	A	0.479	0.360	0.004	0.439	0.276	0.467	0.728
rs2723264	T	0.173	0.205	0.319	0.190	0.541	0.185	0.641
rs17465912	T	0.151	0.182	0.302	0.153	0.952	0.147	0.851
rs10784462	C	0.481	0.511	0.453	0.500	0.590	0.472	0.792
rs11175784	C	0.509	0.517	0.528	0.517	0.478	0.496	0.878
rs10784470	T	0.357	0.301	0.149	0.339	0.589	0.350	0.812
rs7308193	G	0.314	0.341	0.481	0.306	0.800	0.329	0.641
rs10506150	T	0.122	0.085	0.163	0.099	0.328	0.129	0.725
rs7309197	T	0.490	0.517	0.512	0.508	0.617	0.500	0.774
rs17491061	C	0.157	0.182	0.401	0.153	0.886	0.140	0.487
rs10784499	T	0.337	0.364	0.493	0.347	0.769	0.325	0.703
rs10878343	T	0.204	0.224	0.544	0.219	0.604	0.227	0.392
rs11564128	T	0.147	0.182	0.240	0.153	0.829	0.147	0.982
rs2896975	G	0.330	0.364	0.375	0.339	0.781	0.346	0.597
rs10467144	A	0.432	0.476	0.282	0.440	0.837	0.446	0.680
rs11175985	T	0.147	0.097	0.075	0.120	0.276	0.147	0.995
rs4768230	A	0.331	0.273	0.124	0.314	0.604	0.325	0.842
rs10878371	T	0.426	0.477	0.207	0.445	0.590	0.447	0.536
rs11176030	T	0.334	0.278	0.146	0.314	0.553	0.325	0.780
rs2404832	G	0.338	0.364	0.509	0.347	0.790	0.329	0.764
rs17444124	C	0.479	0.455	0.540	0.467	0.725	0.476	0.906
rs6581667	G	0.470	0.466	0.914	0.446	0.500	0.476	0.876
rs10506155	A	0.333	0.310	0.562	0.322	0.758	0.339	0.837
rs2162471	C	0.178	0.188	0.758	0.193	0.578	0.206	0.286
rs10459265	G	0.472	0.466	0.877	0.446	0.466	0.476	0.921
rs4767973	A	0.311	0.347	0.352	0.310	0.967	0.329	0.575
rs11835105	G	0.803	0.227	0.360	0.215	0.539	0.168	0.261
rs17466605	A	0.313	0.347	0.379	0.306	0.822	0.336	0.470
rs17520676	-	0	0		0		0	

Table 8-6 LRRK2 CD sub phenotypic analysis: disease location, p-values<0.05 highlighted

SNP	Minor allele	Control MAF	B1 MAF	p-value	B2 MAF	p-value	B3 MAF	p-value
rs7973254	C	0.315	0.294	0.419	0.316	0.991	0.385	0.149
rs12230685	T	0.119	0.103	0.373	0.149	0.347	0.231	0.001
rs10878246	G	0.185	0.193	0.704	0.228	0.263	0.194	0.808
rs6581622	C	0.265	0.262	0.886	0.281	0.725	0.370	0.020
rs10878258	G	0.264	0.245	0.418	0.333	0.116	0.361	0.032
rs7955902	A	0.479	0.473	0.834	0.339	0.005	0.373	0.041
rs2723264	T	0.173	0.167	0.773	0.219	0.225	0.287	0.004
rs17465912	T	0.151	0.148	0.870	0.184	0.357	0.157	0.867
rs10784462	C	0.481	0.470	0.697	0.500	0.697	0.565	0.096
rs11175784	C	0.509	0.485	0.815	0.518	0.595	0.407	0.045
rs10784470	T	0.357	0.369	0.659	0.289	0.150	0.250	0.026
rs7308193	G	0.314	0.298	0.537	0.360	0.323	0.417	0.030
rs10506150	T	0.122	0.116	0.750	0.114	0.813	0.065	0.080
rs7309197	T	0.490	0.483	0.786	0.526	0.467	0.585	0.064
rs17491061	C	0.157	0.144	0.524	0.184	0.446	0.852	0.818
rs10784499	T	0.337	0.330	0.799	0.351	0.769	0.333	0.936
rs10878343	T	0.204	0.208	0.854	0.246	0.299	0.292	0.034
rs11564128	T	0.147	0.146	0.940	0.184	0.298	0.157	0.781
rs2896975	G	0.330	0.330	0.971	0.377	0.305	0.417	0.068
rs10467144	A	0.432	0.429	0.897	0.500	0.170	0.510	0.130
rs11175985	T	0.147	0.131	0.408	0.158	0.756	0.065	0.019
rs4768230	A	0.331	0.339	0.771	0.263	0.140	0.250	0.085
rs10878371	T	0.426	0.433	0.817	0.500	0.135	0.509	0.098
rs11176030	T	0.334	0.341	0.784	0.263	0.126	0.250	0.076
rs2404832	G	0.338	0.330	0.771	0.360	0.645	0.333	0.920
rs17444124	C	0.479	0.487	0.785	0.447	0.515	0.407	0.154
rs6581667	G	0.470	0.444	0.347	0.482	0.805	0.528	0.255
rs10506155	A	0.333	0.313	0.459	0.333	0.989	0.387	0.262
rs2162471	C	0.178	0.179	0.957	0.219	0.277	0.278	0.011
rs10459265	G	0.472	0.444	0.313	0.482	0.835	0.528	0.271
rs4767973	A	0.311	0.305	0.798	0.360	0.292	0.407	0.042
rs11835105	G	0.803	0.189	0.701	0.211	0.737	0.176	0.594
rs17466605	A	0.313	0.307	0.806	0.360	0.312	0.407	0.046
rs17520676	-	0	0		0		0	

Table 8-7 LRRK2 CD sub phenotypic analysis: disease behaviour, p-values<0.05 highlighted

	SNP	Minor allele	Control MAF	E3 MAF	p-value	E2 MAF	p val	E1 MAF	p-value
1	rs7973254	C	0.315	0.274	0.1977	0.395	0.044	0.444	0.1023
2	rs12230685	T	0.119	0.121	0.908	0.134	0.5745	0.105	0.8
3	rs10878246	G	0.185	0.246	0.0234	0.25	0.0497	0.263	0.2246
4	rs6581622	C	0.265	0.25	0.6087	0.305	0.2892	0.289	0.7408
5	rs10878258	G	0.264	0.243	0.4669	0.287	0.5501	0.237	0.7051
6	rs7955902	A	0.479	0.42	0.0926	0.393	0.0512	0.389	0.2897
7	rs2723264	T	0.173	0.165	0.7554	0.189	0.6263	0.105	0.2729
8	rs17465912	T	0.151	0.206	0.0299	0.171	0.5224	0.184	0.5798
9	rs10784462	C	0.481	0.441	0.2437	0.482	0.9829	0.5	0.8161
10	rs11175784	C	0.509	0.452	0.2504	0.488	0.9307	0.5	0.9164
11	rs10784470	T	0.357	0.337	0.5356	0.335	0.5857	0.316	0.5998
12	rs7308193	G	0.314	0.294	0.5241	0.305	0.8113	0.289	0.7471
13	rs10506150	T	0.122	0.132	0.6331	0.116	0.8324	0.053	0.1972
14	rs7309197	T	0.49	0.452	0.2616	0.488	0.9507	0.5	0.9075
15	rs17491061	C	0.157	0.204	0.0656	0.173	0.6006	0.184	0.647
16	rs10784499	T	0.337	0.382	0.163	0.36	0.5703	0.395	0.4615
17	rs10878343	T	0.204	0.195	0.7373	0.213	0.7819	0.132	0.2741
18	rs11564128	T	0.147	0.206	0.019	0.171	0.4375	0.184	0.531
19	rs2896975	G	0.33	0.294	0.266	0.329	0.9953	0.342	0.8711
20	rs10467144	A	0.432	0.383	0.1456	0.468	0.4103	0.472	0.6341
21	rs11175985	T	0.147	0.169	0.3655	0.134	0.6642	0.132	0.7917
22	rs4768230	A	0.331	0.324	0.8053	0.311	0.6044	0.263	0.379
23	rs10878371	T	0.426	0.386	0.2341	0.469	0.3055	0.474	0.561
24	rs11176030	T	0.334	0.324	0.7446	0.311	0.5607	0.263	0.3624
25	rs2404832	G	0.338	0.379	0.2106	0.36	0.5869	0.395	0.4693
26	rs17444124	C	0.479	0.529	0.1432	0.482	0.958	0.447	0.6971
27	rs6581667	G	0.47	0.401	0.0402	0.512	0.318	0.474	0.9673
28	rs10506155	A	0.333	0.294	0.2267	0.384	0.1961	0.237	0.2169
29	rs2162471	C	0.178	0.179	0.9624	0.183	0.8753	0.105	0.2478
30	rs10459265	G	0.472	0.401	0.035	0.512	0.3407	0.474	0.9859
31	rs4767973	A	0.311	0.276	0.256	0.311	0.9933	0.316	0.9532
32	rs11835105	G	0.803	0.221	0.3947	0.207	0.7654	0.237	0.5487
33	rs17466605	A	0.313	0.276	0.2319	0.317	0.9212	0.316	0.9732
34	rs17520676	-	0	0		0		0	

Table 8-8 LRRK2 UC sub phenotypic analysis, p-values<0.05 highlighted

8.4.3 Haplotype analysis

When corrected for multiple testing across the 43 haplotypes, a significant p-value was defined as being <0.00012.

8.4.3.1 IBD, CD and UC overall

Results of a haplotype analysis in CD and UC are shown in Table 8-9 and Table 8-10. Whilst none of the haplotypes reached statistical significance when corrected for multiple testing, there were haplotypes at the 5' end of the gene (blocks 1 and 2) that approached significance. In block 1, the TCT and TCG haplotypes were associated with UC ($p = 0.02$ and $p = 0.013$, respectively). In block 2, the TAA and TAC haplotypes were associated with both IBD and UC (TAA, IBD $p = 0.018$, UC $p = 0.039$) and (TAC, IBD $p = 0.0224$, UC $p = 0.0165$). None of the haplotypes were associated with CD

Haplotype	Control	IBD freq	IBD p-value	CD freq	CD p-value	UC freq	UC p-value
Block 1							
TCT	0.508	0.474	0.0924	0.494	0.5859	0.446	0.0201
TCG	0.176	0.207	0.0506	0.193	0.3867	0.228	0.0134
CCT	0.189	0.18	0.5771	0.174	0.4028	0.181	0.7559
CTT	0.118	0.122	0.7721	0.129	0.5142	0.12	0.958
CCG	0.009	0.016	0.1207			0.025	0.0157
Block 2							
TAA	0.46	0.412	0.0178	0.42	0.0911	0.404	0.0393
TAC	0.23	0.271	0.0224	0.254	0.236	0.286	0.0165
CGC	0.208	0.218	0.5745	0.227	0.3385	0.205	0.9218
CAC	0.043	0.053	0.2275	0.051	0.3962	0.057	0.1881
TGC	0.045	0.043	0.8157	0.045	0.9871	0.041	0.7337
Block 3							
CC	0.677	0.654	0.2393	0.651	0.2521	0.641	0.1516
TC	0.172	0.176	0.816	0.19	0.329	0.168	0.8662
CT	0.15	0.17	0.1857	0.158	0.6109	0.189	0.0432
Block 4							
G	0.519	0.52	0.9711	0.514	0.8212	0.542	0.386
C	0.481	0.48	0.9711	0.486	0.8212	0.458	0.386
Block 5							
T	0.509	0.507	0.9531	0.497	0.638	0.533	0.3669
C	0.491	0.493	0.9531	0.503	0.638	0.467	0.3669
Block 6							
GC	0.331	0.363	0.1066	0.343	0.5977	0.374	0.0961
TC	0.354	0.329	0.1915	0.334	0.3686	0.331	0.354
GG	0.312	0.303	0.647	0.319	0.7563	0.291	0.3833
Block 7							
A	0.878	0.888	0.4772	0.891	0.3962	0.878	0.9988
T	0.122	0.112	0.4772	0.109	0.3962	0.122	0.9988
Block 8							
TT	0.49	0.493	0.8812	0.502	0.6426	0.469	0.4281
AT	0.354	0.338	0.4244	0.343	0.6297	0.34	0.5975
AC	0.156	0.168	0.4084	0.155	0.9955	0.19	0.0825
Block 9							
CC	0.459	0.432	0.1833	0.437	0.3707	0.423	0.1775
TC	0.337	0.363	0.1954	0.339	0.926	0.376	0.1255
CT	0.204	0.206	0.9156	0.223	0.326	0.201	0.8805

Table 8-9 LRRK2 Haplotype analysis in IBD, CD and UC, haplotype blocks 1-9, p-values<0.05 highlighted

Haplotype	Control	IBD freq	IBD p-value	CD freq	CD p-value	UC freq	UC p-value
Block 10							
CGACG	0.33	0.326	0.8441	0.346	0.4654	0.311	0.4572
CTGCA	0.183	0.174	0.5688	0.187	0.8536	0.158	0.2233
TTGCG	0.148	0.17	0.1565	0.158	0.5817	0.191	0.0315
CTGTA	0.148	0.137	0.423	0.126	0.1943	0.154	0.7472
CTACG	0.104	0.113	0.4809	0.105	0.9578	0.115	0.493
CTGCG	0.087	0.081	0.5872	0.078	0.5115	0.07	0.2552
Block 11							
TAAA	0.326	0.324	0.9191	0.343	0.4385	0.311	0.5288
CTAC	0.332	0.311	0.2728	0.313	0.4158	0.312	0.4565
CAGC	0.152	0.171	0.2126	0.158	0.7323	0.191	0.0458
TAGA	0.102	0.113	0.409	0.105	0.8223	0.114	0.4924
CAGA	0.084	0.08	0.7028	0.077	0.5757	0.069	0.318
Block 12							
CGTA	0.53	0.544	0.4786	0.54	0.6138	0.556	0.2925
GACG	0.176	0.181	0.7412	0.195	0.3104	0.178	0.9323
GATG	0.156	0.142	0.3065	0.131	0.137	0.147	0.5921
GGTG	0.138	0.133	0.7329	0.134	0.8165	0.12	0.3311
Block 13							
GTGA	0.49	0.48	0.6146	0.474	0.5189	0.486	0.8841
ATAA	0.311	0.309	0.9261	0.324	0.5378	0.29	0.3875
GGGA	0.196	0.209	0.4401	0.198	0.9402	0.222	0.2264

Table 8-10 LRRK2 Haplotype analysis in IBD, CD and UC, haplotype blocks 10-13, p-values<0.05 highlighted

8.4.3.2 CD sub phenotype

An analysis on the CD phenotypic data is shown in Table 8-11/Table 8-12 (disease location) and Table 8-13/Table 8-14 (disease behaviour). There were 3 haplotypes that demonstrated a non significant trend to association with L1 disease location, the strongest p-value being 0.015 for TAA in block 2. When disease behaviour was analysed the results mirrored the single SNP analysis. A few haplotypes showed a trend to association with B3 disease: the CTT haplotype in block 1 (p-value 0.0009, haplotype frequency 0.23 in patients with B3 disease, 0.12 in controls). A further 10 haplotypes across the gene also showed an association which approached statistical significance (p-value range 0.0036-0.0408).

Haplotype	Control freq	L1 freq	p-value	L2 freq	p-value	L3 freq	p-value
Block 1							
TCT	0.506	0.417	0.0297	0.518	0.7483	0.522	0.661
CCT	0.19	0.208	0.5919	0.176	0.6254	0.157	0.2049
TCG	0.178	0.213	0.2645	0.185	0.7741	0.181	0.8707
CTT	0.119	0.153	0.1968	0.112	0.7538	0.129	0.6264
CCG	0.008						
Block 2							
TAA	0.459	0.361	0.0149	0.418	0.2396	0.453	0.8308
TAC	0.231	0.285	0.1179	0.253	0.4433	0.23	0.9795
CGC	0.208	0.274	0.049	0.213	0.8494	0.211	0.9048
TGC	0.045	0.036	0.608	0.052	0.6382	0.047	0.8585
CAC	0.041	0.041	0.9972	0.059	0.2524	0.053	0.4177
CGA	0.012	0.002	0.2492	0.004	0.2495	0.004	0.2394
Block 3							
CC	0.677	0.616	0.1088	0.659	0.5803	0.669	0.787
TC	0.171	0.203	0.317	0.188	0.5316	0.185	0.6096
CT	0.149	0.18	0.2969	0.151	0.9358	0.146	0.8915
Block 4							
G	0.519	0.489	0.4532	0.5	0.5905	0.528	0.7924
C	0.481	0.511	0.4532	0.5	0.5905	0.472	0.7924
Block 5							
T	0.509	0.483	0.5282	0.484	0.4815	0.504	0.8782
C	0.491	0.517	0.5282	0.516	0.4815	0.496	0.8782
Block 6							
TC	0.356	0.301	0.1583	0.339	0.6174	0.345	0.758
GC	0.33	0.358	0.4663	0.356	0.4504	0.326	0.8817
GG	0.313	0.341	0.464	0.305	0.8244	0.324	0.6987
Block 7							
A	0.878	0.915	0.1634	0.901	0.3278	0.871	0.7249
T	0.122	0.085	0.1634	0.099	0.3278	0.129	0.7249
Block 8							
TT	0.49	0.518	0.4975	0.508	0.6166	0.5	0.7737
AT	0.354	0.3	0.1637	0.339	0.6499	0.36	0.8582
AC	0.155	0.182	0.3767	0.153	0.926	0.14	0.5228
Block 9							
CC	0.459	0.411	0.2404	0.434	0.4828	0.448	0.7348
TC	0.337	0.364	0.4931	0.347	0.7685	0.325	0.7034
CT	0.204	0.225	0.522	0.219	0.6041	0.227	0.3916

Table 8-11 LRRK2 CD disease location haplotype analysis, haplotypes 1-9, p-values<0.05 highlighted.

Haplotype	Control freq	L1 freq	p-value	L2 freq	p-value	L3 freq	p-value
Block 10							
CGACG	0.33	0.364	0.3746	0.339	0.7809	0.346	0.5965
CTGCA	0.183	0.176	0.8194	0.194	0.6944	0.178	0.8483
TTGCG	0.148	0.182	0.2549	0.153	0.8586	0.147	0.9489
CTGTA	0.148	0.097	0.0691	0.12	0.2575	0.147	0.9557
CTACG	0.104	0.111	0.7686	0.106	0.9322	0.101	0.9091
CTGCG	0.087	0.071	0.4749	0.089	0.9349	0.08	0.7249
Block 11							
TAAA	0.326	0.358	0.4107	0.339	0.7103	0.325	0.8395
CTAC	0.332	0.272	0.1186	0.314	0.5865	0.343	0.5894
CAGC	0.152	0.182	0.3026	0.153	0.9451	0.147	0.8326
TAGA	0.102	0.113	0.6587	0.107	0.8245	0.101	0.9635
CAGA	0.084	0.068	0.4661	0.087	0.9173	0.081	0.8235
Block 12							
CGTA	0.528	0.534	0.8767	0.554	0.4661	0.528	0.9907
GACG	0.176	0.188	0.7125	0.195	0.4989	0.199	0.3711
GATG	0.157	0.125	0.2798	0.128	0.2545	0.139	0.4562
GGTG	0.137	0.153	0.5766	0.124	0.5823	0.134	0.8766
Block 13							
GTGA	0.49	0.426	0.1145	0.479	0.7765	0.497	0.855
ATAA	0.311	0.347	0.3468	0.306	0.8901	0.329	0.5674
GGGA	0.196	0.227	0.3343	0.211	0.6136	0.168	0.2857

Table 8-12 LRRK2 CD disease location haplotype analysis, haplotypes 10-13, p-values<0.05 highlighted.

Haplotype	Control	B1 freq	p-value	B2 freq	p-value	B3 freq	p val
Block 1							
TCT	0.506	0.524	0.5616	0.46	0.3517	0.415	0.0723
CCT	0.19	0.18	0.6896	0.163	0.4733	0.159	0.4254
TCG	0.178	0.179	0.9038	0.224	0.2257	0.193	0.6963
CTT	0.119	0.103	0.3729	0.149	0.3468	0.231	9.00E-04
CCG	0.008	0.014	0.3101				
Block 2							
TAA	0.459	0.458	0.9389	0.33	0.0086	0.354	0.0361
TAC	0.231	0.239	0.6662	0.316	0.0424	0.226	0.9291
CGC	0.208	0.201	0.7637	0.257	0.2203	0.308	0.0158
TGC	0.045	0.041	0.6929	0.073	0.1823	0.049	0.8442
CAC	0.041	0.057	0.2231	0.02	0.2594	0.056	0.4652
CGA	0.012			0.004	0.4128	0.004	0.4544
Block 3							
CC	0.677	0.686	0.7442	0.598	0.089	0.557	0.0121
TC	0.171	0.166	0.8034	0.218	0.2203	0.285	0.0036
CT	0.149	0.147	0.9081	0.182	0.348	0.156	0.8593
Block 4							
G	0.519	0.53	0.6968	0.5	0.697	0.435	0.0964
C	0.481	0.47	0.6968	0.5	0.697	0.565	0.0964
Block 5							
T	0.509	0.515	0.8161	0.482	0.5952	0.593	0.0452
C	0.491	0.485	0.8161	0.518	0.5952	0.407	0.0452
Block 6							
TC	0.356	0.366	0.6806	0.289	0.1571	0.25	0.0277
GC	0.33	0.335	0.852	0.351	0.6505	0.334	0.9445
GG	0.313	0.296	0.5085	0.359	0.3122	0.416	0.0283
Block 7							
A	0.878	0.884	0.7499	0.886	0.8129	0.935	0.0795
T	0.122	0.116	0.7499	0.114	0.8129	0.065	0.0795
Block 8							
TT	0.49	0.483	0.7858	0.526	0.4665	0.585	0.0612
AT	0.354	0.373	0.4798	0.289	0.1684	0.267	0.0691
AC	0.155	0.144	0.5706	0.184	0.4238	0.148	0.8439
Block 9							
CC	0.459	0.461	0.9265	0.404	0.2601	0.374	0.0906
TC	0.337	0.33	0.7991	0.351	0.7689	0.333	0.9361
CT	0.204	0.208	0.8543	0.246	0.299	0.293	0.0314

Table 8-13 LRRK2 CD behaviour haplotype analysis, haplotypes 1-9, p-values<0.05 highlighted

Haplotype	Control	B1 freq	p-value	B2 freq	p-value	B3 freq	p val
Block 10							
CGACG	0.33	0.33	0.9705	0.377	0.3054	0.417	0.0682
CTGCA	0.183	0.208	0.2577	0.105	0.0378	0.185	0.9623
TTGCG	0.148	0.146	0.9015	0.184	0.3119	0.157	0.8019
CTGTA	0.148	0.131	0.3789	0.158	0.7834	0.065	0.0177
CTACG	0.104	0.101	0.8753	0.123	0.5304	0.093	0.7175
CTGCG	0.087	0.083	0.8222	0.053	0.2094	0.083	0.8969
Block 11							
TAAA	0.326	0.326	0.9867	0.377	0.2741	0.417	0.0585
CTAC	0.332	0.339	0.7769	0.263	0.1356	0.25	0.0835
CAGC	0.152	0.146	0.7665	0.185	0.3584	0.158	0.8717
TAGA	0.102	0.103	0.9718	0.122	0.5023	0.092	0.7525
CAGA	0.084	0.082	0.8469	0.053	0.2379	0.083	0.9608
Block 12							
CGTA	0.528	0.556	0.3132	0.518	0.8354	0.472	0.2712
GACG	0.176	0.179	0.8727	0.21	0.3607	0.278	0.0097
GATG	0.157	0.134	0.2467	0.123	0.3368	0.112	0.2132
GGTG	0.137	0.131	0.7339	0.14	0.928	0.138	0.9764
Block 13							
GTGA	0.49	0.504	0.5945	0.43	0.2192	0.417	0.1444
ATAA	0.311	0.303	0.7626	0.36	0.2881	0.407	0.0408
GGGA	0.196	0.187	0.6605	0.211	0.7071	0.176	0.6194

Table 8-14 LRRK2 CD disease behaviour haplotype analysis, haplotypes 10-13, p-values<0.05 highlighted

8.4.3.3 UC sub phenotype

An analysis of the UC phenotypic data is shown in Table 8-15. The CCG haplotype in block 1, which showed a statistically significant association with disease extent, also had very low carriage rates (carriage rates: E3 was not calculated, E2 0.043 and E1 0.063); therefore the levels of significance are uncertain (E2 p=0.0008; E1 p=0.001). A number of other haplotypes showed a trend to association with disease extent in blocks 1 and 2 but these did not reach overall statistical significance: in block 1, the TCT haplotype with E2 disease extent (p=0.0067) and the TCG haplotype with E3 disease (p=0.0156); in block 2, the TAC haplotype with E3 disease extent (p=0.0397). Three other haplotypes showed a trend to association with E3 disease extent: TTGCG in block 10 (p=0.0214), CAGC in block 11 (p=0.0298) and CGTA in block 12 (p=0.035).

Haplotype	Control freq	E3 freq	p-value	E2 freq	p-value	E1 freq	p-value
Block 1							
TCT	0.506	0.48	0.4488	0.395	0.0067	0.345	0.0494
TCG	0.178	0.243	0.0156	0.221	0.3137	0.287	0.1345
CCT	0.191	0.153	0.1464	0.207	0.322	0.2	0.7078
CTT	0.119	0.121	0.908	0.134	0.5745	0.105	0.8
CCG	0.008			0.043	8.00E-04	0.063	0.001
Block 2							
TAA	0.459	0.412	0.1699	0.38	0.0584	0.381	0.3388
TAC	0.231	0.291	0.0397	0.27	0.264	0.3	0.3183
CGC	0.207	0.19	0.5294	0.238	0.3668	0.203	0.9552
TGC	0.045	0.047	0.9031	0.045	0.9857	0.03	0.6569
CAC	0.042	0.052	0.4568	0.061	0.2698	0.08	0.2531
CGA	0.012	0.006	0.3844	0.004	0.3321	0.004	0.6311
Block 3							
CC	0.677	0.632	0.1534	0.641	0.361	0.711	0.6677
TC	0.171	0.162	0.7314	0.188	0.6015	0.105	0.2863
CT	0.149	0.203	0.0307	0.17	0.4941	0.184	0.5508
Block 4							
G	0.519	0.559	0.2438	0.518	0.9829	0.5	0.8161
C	0.481	0.441	0.2438	0.482	0.9829	0.5	0.8161
Block 5							
T	0.509	0.548	0.2507	0.512	0.931	0.5	0.9164
C	0.491	0.452	0.2507	0.488	0.931	0.5	0.9164
Block 6							
TC	0.355	0.33	0.4454	0.335	0.6067	0.315	0.611
GC	0.331	0.376	0.1646	0.36	0.4496	0.395	0.4044
GG	0.312	0.286	0.4039	0.304	0.8292	0.289	0.7567
Block 7							
A	0.878	0.868	0.6331	0.884	0.8324	0.947	0.1972
T	0.122	0.132	0.6331	0.116	0.8324	0.053	0.1972
Block 8							
TT	0.49	0.452	0.2616	0.488	0.9507	0.5	0.9075
AT	0.354	0.343	0.7395	0.341	0.7499	0.316	0.6257
AC	0.156	0.205	0.0529	0.171	0.615	0.184	0.6299
Block 9							
CC	0.459	0.423	0.2877	0.427	0.4444	0.474	0.8566
TC	0.337	0.382	0.163	0.36	0.5703	0.395	0.4615
CT	0.204	0.195	0.7373	0.213	0.7819	0.132	0.2741

Table 8-15 LRRK2 UC disease extent haplotype analysis, haplotypes 1-9, p-values<0.05 highlighted

Haplotype	Control freq	E3 freq	p-value	E2 freq	p-value	E1 freq	p-value
Block 10							
CGACG	0.33	0.294	0.266	0.329	0.9953	0.342	0.8711
CTGCA	0.183	0.154	0.2687	0.177	0.8427	0.132	0.4169
TTGCG	0.148	0.206	0.0214	0.171	0.4587	0.184	0.543
CTGTA	0.148	0.169	0.394	0.134	0.637	0.132	0.777
CTACG	0.104	0.092	0.5765	0.14	0.1642	0.146	0.4037
CTGCG	0.087	0.084	0.887	0.049	0.0983	0.064	0.6275
Block 11							
CTAC	0.331	0.323	0.8062	0.311	0.5909	0.262	0.3799
TAAA	0.327	0.294	0.3061	0.329	0.9418	0.342	0.8353
CAGC	0.152	0.206	0.0298	0.171	0.5239	0.185	0.5796
TAGA	0.102	0.092	0.6155	0.14	0.1498	0.131	0.5617
CAGA	0.084	0.081	0.8529	0.049	0.1179	0.079	0.9055
Block 12							
CGTA	0.528	0.599	0.035	0.488	0.3407	0.526	0.9859
GACG	0.176	0.18	0.8652	0.183	0.8265	0.105	0.2584
GATG	0.157	0.114	0.0737	0.201	0.1549	0.132	0.6717
GGTG	0.137	0.107	0.1807	0.128	0.7474	0.237	0.0834
Block 13							
GTGA	0.49	0.504	0.6966	0.476	0.7246	0.447	0.6022
ATAA	0.311	0.276	0.2608	0.311	0.9993	0.316	0.9494
GGGA	0.196	0.221	0.3631	0.207	0.73	0.237	0.532

Table 8-16 LRRK2 UC disease extent haplotype analysis, haplotypes 10-13 p-values<0.05 highlighted

8.5 Discussion

Mucins make attractive potential candidate genes in IBD susceptibility, as they are important constituents of the mucus layer and as such, are potentially involved in barrier function in the gut. However, for MUC19 there are no data in the literature demonstrating *MUC19* expression in the GI tract or in immune cells. The *MUC19* data presented in this chapter had sufficient power to confidently rule out any potential genotypic association between IBD and this gene. In addition, looking across all known SNPs in the *MUC19* gene on available HapMap data, only one SNP has any LD with rs11175593. Thus it is unlikely that the candidate gene for the rs11175593 locus is *MUC19*.

Multiple SNPs within the *LRRK2* gene, have D' and r^2 of 1 with rs11175593, thus it is more likely that *LRRK2* represents the susceptibility gene for the rs11175593 signal at this region. The complicated haplotypic structure and large gene size of *LRRK2* makes it more difficult to study than the smaller *MUC19* gene. This has been borne out with the *LRRK2* results presented here, which showed an association, the significance of which was lost when corrected for multiple testing. However, the studied population of 639 patients (336 CD) and 539 controls had minimal power to detect differences, with a possibility of a type 2 statistical error. For a single SNP analysis to have >80% power to detect a difference with an OR of 1.5, at least 384 patients are required. With correction for multiple testing, the number needed increases to at least 836 patients. Therefore, these *LRRK2* studies were inadequately powered to rule out an association between *LRRK2* and the SNP rs11175593. Given the fact that the *MUC19* study was negative, and there was a high OR of 1.5 for rs11175593 in the meta-analysis, it is likely that the some of the SNP results are false negatives. The SNPs that showed an association before multiple testing at least should be further investigated in a larger population.

In this chapter, the results demonstrated that there are more potentially significant associations between *LRRK2* and susceptibility to UC than to CD. This is especially interesting as the SNP was associated with CD susceptibility in the CD meta-analysis; in addition, other autophagy genes (*ATG16L1* and *IRGM*) are also associated with CD much more so than with UC. UC GWAS have not found an association between UC and *LRRK2*.

Intriguingly, the most significant association was between B3 disease in CD and a haplotype (CTT in haplotype block 1). There were other *LRRK2* SNPs and haplotypes associations with CD, even though significance was lost in CD when corrected for multiple testing. This suggests that the gene may be involved in determining disease phenotype rather than susceptibility to CD. This demonstrates why sub phenotypic studies are valuable, even with an overall lack of association of a gene with disease susceptibility.

In order to fully understand the role of *LRRK2* in CD susceptibility in the Scottish population, a larger cohort is required to provide more power. In particular, SNPs at the 5’ end of the gene, especially in the first 2 haplotype blocks, should be genotyped, followed by replication in a further cohort. In addition, exonic sequencing of the *LRRK2* gene would be helpful to attempt to find a non synonymous SNP that may be associated with disease susceptibility.

Functional studies of autophagy proteins (ATG16L1 and IRGM) associated with CD have suggested that mutations decrease autophagy[103;323], resulting in defective intracellular processing of bacteria. Thus, if LRRK2 is implicated in CD pathogenesis and functionally involved in autophagy, CD-associated *LRRK2* mutations would be expected to reduce autophagy. In PD, the *LRRK2* mutations prevent the negative regulation of autophagy, thus increasing autophagy. Certainly, the PD autosomal dominant (AD) mutations are unlikely to be associated with CD as there is no known link between CD susceptibility and PD. Alternatively, LRRK2 pathogenicity in CD could be mediated by IFN-γ and its downstream effects, as argued by Gardet *et al*.[322] Further functional studies to delineate the pathways involving LRRK2 are required.

Chapter 9 Future work

9.1 Crohn's Disease Phenotype

The studies presented indicate that the risk of disease progression in the Scottish cohort studied is lower than in other cohorts, yet the risk of resection is higher. It consolidates existing knowledge that disease location is the most important clinical determinant of disease progression and the need for surgery. The risk of further surgery also appears to be determined by the disease location, with colonic disease behaving in a more benign fashion than patients with ileal and/or UGI disease. Another important study would be to examine the risk of disease progression and surgery with respect to medication. With increasing use of thiopurines and biologicals, conclusive evidence is still lacking that these medications actually change disease course.[324] It would be important to factor these variables into a multivariate analysis. An alternative way of considering this, possible with the currently available data, would be to complete an analysis of disease progression with respect to decade at diagnosis. This could provide circumstantial evidence that immunosuppressants are changing the CD course, as thiopurines have only been used widely since the early 1990s. It would also be useful to repeat the multivariate analysis of time to surgery including decade at diagnosis in the Cox proportional hazards model, as it would appear to be important, reflecting a change in surgical practices over time, and possibly reflecting a change in use of steroid-sparing agents – but this would be best done with the additional knowledge of medication history, which is not available in a substantial proportion of the Scottish cohort. This could help to tease out whether surgical practices have changed, or whether steroid sparing agents are actually changing disease course.

The analysis of patients with stricturing and penetrating disease indicates that the B2 and B3 categories should be considered as completely separate disease progression categories, something that has implications for future CD classification. This study should be extended to include all Scottish patients rather than just the Dundee subset which could provide a more robust confirmation of the findings presented. In addition, a prospective study involving the pathologists who report the surgical resection specimens would be necessary to ensure that subtle details of the specimen are included that might not be documented on a standard pathology report but which might affect the decision as to whether there are strictures or fistulae present in the pathology specimen.

Of course, in the future, prospective studies of CD natural history that are population based will be important, although this will take some time to accumulate the relevant data. In countries like the UK where there are no centralised hospital record systems, it can be hard to ensure that the entire disease population of an area has been accounted for. Therefore clinic-based studies are the most realistic mode of recruitment for prospective studies in the UK for the time being. Careful thought would need to be given as to which parameters should be collected from these patients. Certainly details of all investigations, medications taken, microbiology, biochemistry and pathology results, disease activity index at predefined points (using the Crohn's Disease Activity Index or Harvey Bradshaw Index) and the Endoscopic Index of Severity and Physician Global Assessment indices should be recorded. In addition, blood for DNA extraction, serum extraction and stool samples should be taken from consenting patients.

9.2 Crohn's disease severity

Predicting disease severity is fertile ground for future studies that could have important implications in CD patient care. The disease severity score is a completely novel way of considering CD severity that has been shown to be correlated with long term disease progression. It has been possible, with 3 clinical factors and 2 genetic factors at diagnosis, to build a model that can predict with reasonable accuracy whether a person is likely to have more severe disease. This has important implications for choosing patients for early top-down therapy.

The most important future work necessary for this study is confirmation of the findings in a separate cohort. This confirmation cohort would need to have their novel severity score calculated. In addition, the 5 variables from which the model has been produced (age <17 at diagnosis, ileal and/or UGI disease at diagnosis, rs13361189 and rs9286879 genotype) would also need to be obtained. A further ROC curve would then be made from the sensitivities and specificities of the modelled probability to predict more severe disease. There are plans to complete this in the Edinburgh cohort in the near future. A separate cohort, geographically separate from Scotland would also be important if this formula is to have more widespread use.

Additionally, it would be interesting to repeat the analyses with a different cut-off point for defining more severe disease. This study has used a cut-off that divides the cohorts into two approximately equal halves. Defining 'more severe' disease as a higher or lower severity score could produce a different set of variables. This is likely in view of the Beaugerie[234] variables which predict 'disabling disease'. These variables are different to the ones suggested in this thesis but this is unsurprising considering their definition encompasses a greater proportion of patients.

A way to develop the formula further would be to use SNPs that appeared to differentiate between B1 and B2/B3 disease on GWAS, as had been initially planned in this study. This could provide more relevant genotyping to enter into a multivariate analysis, and could increase the sensitivity and specificity of the model to predict more severe disease. Any of the SNPs thus uncovered would be worth further detailed investigation as they might help to pin-point the molecular pathways involved in disease evolution – something that would be useful for future drug development. Immune responses to microbial antigens have previously been shown to be associated with CD progression (e.g. anti-Saccharomyces antibodies[19], anti-OmpC[325] and anti-I2 antibodies[325]); they are also likely to be correlated with the novel severity score. All these factors could be considered for inclusion in the model with the aim of increasing the accuracy of the model for disease prediction.

In schizophrenia, polygenic scoring across schizophrenia GWAS SNPs with p-values of <0.5 is positively associated with disease susceptibility.[237] It would be interesting to do this with the WTCCC GWAS data and calculate the polygenic score, and compare CD patients with controls. To take this one step further, using the polygenic score to look for differences between more and less severe patients would also be worthwhile, although this would require severity score calculation in all the WTCCC patients. Whilst polygenic scoring would not have implications for patient care at the moment, in the future it could do so as genotyping costs come down.

9.3 GALNT2 genotype

Any future genotyping work using the Illumina® platform in the Gastrointestinal Unit will benefit from lessons learnt from the *GALNT2* genotyping work. Indeed, it has highlighted the problems with automated genotyping, and demonstrates the need for stringent quality control as well as the importance of replication of any positive result.

It has been demonstrated that rs7536663 in the *GALNT2* gene is associated with CD susceptibility in the Edinburgh cohort with an OR of 1.38. Although the MAF in CD and controls in the Dundee replication cohort were different and the differences in a similar direction to the Edinburgh cohort MAF, it was not statistically significant. The rs7536663 SNP and its associated haplotype should be genotyped in a larger cohort. The rs7536663 SNP is currently being genotyped on the WTCCC immunochip with hopefully the results being available shortly. Should this provide robust replication of the Edinburgh cohort findings, a more detailed examination of the *GALNT2* gene should be completed, including further sequencing across the exons and important regulatory regions. As the rs7536663 SNP lies within intron 1, this search should concentrate on exons 1 and 2, and the regulatory regions around it due to the large size of the gene. Next-generation sequencing is making cost-effective whole gene sequencing a reality thus it is more likely to represent the best way forward in discovering a mutation that has direct implications in the biological pathways of CD.

9.4 GALNT2 expression

It has been demonstrated convincingly that NOD2 and GALNT2 interact in mammalian cells. This validates the initial yeast two-hybrid experiment that showed their interaction in yeast cells. However the CoIP experiments need to be repeated to ensure that these results can be replicated. It would be good to do this with densitometry on the western blot bands to quantitate the differing band intensities between the NOD2 mutants.

Most important to the further development of the experiments in this chapter would be the development of a NOD2 antibody that works for both IHC and western blotting. In an ideal world, if the NOD2 protein could be produced in sufficient quantities it would be the preferred immunogen. Failing that, further peptide sequences in the OMAP formation could be a better type of immunogen than KLH and might prove successful in antibody production. Once a working NOD2 antibody is produced, IHC and immunofluorescence should be used on colonic and ileal tissues, to investigate the protein co-expression of NOD2 and GALNT2. Performing cell fractionation prior to CoIP could also help to prove that they are expressed in the same subcellular compartments.

For the studies on mRNA expression, it would be useful to look at *NOD2* and *GALNT2* expression in tissue samples, including PBMCs as well as gastrointestinal tract tissue. Microdissection of different cell types with subsequent qPCR could provide evidence of cellular co-localization of NOD2 and GALNT2.

9.5 MUC2 and MUC3A genotyping

In the first instance, the *MUC2* SNP rs7942850 and its two-marker haplotype need full and formal replication in a large cohort that contains controls. With *MUC3A*, a more detailed tagging SNP analysis, tagging for a lower haplotype frequency would be useful. This was not completed in the study presented here due to the cost implications. Improving knowledge of the full sequence of *MUC3A* would also help to provide an adequately genotyped study. Interest in MUC3A was on the basis of a study using RFLP markers.[286] However, the RLFP marker that was implicated in IBD susceptibility is actually >20Mbp upstream of the gene. More recent UC GWAS has shown that a marker nearer this RFLP (rs7807999) is associated with UC susceptibility. This SNP is in between the *SMURF1* and *KPNA7* genes. Therefore further tagging SNP genotyping studies across these genes should also be completed.

Mucins may be implicated in IBD pathogenesis without requiring germline variation in the underlying genes. As already discussed, *MUC2*[305]and *MUC3A*[306] expression in colon cancer cell lines has been shown to be highly controlled by methylation and histone deacetylation. Therefore epigenetic changes could affect the quantity of mucin produced. Assessing human colonic tissue DNA methylation with bisulfite sequencing of extracted DNA with assessment of *MUC2* and *MUC3A* mRNA expression in the same tissues could be completed. Given the literature suggesting that *MUC2* expression is partly regulated by short-chain fatty acids (fermentation products of bacteria) via epigenetics[307], it would be worthwhile to assess the corresponding bacteria present in the faeces and mucus using 16S rRNA technology. Further bacterial studies examining in more detail the effects of specific bacteria – for example adherent-invasive *E. Coli* – on mucin expression and production in cell lines would be useful.

9.6 MUC19 and LRRK2 genotyping

The results presented in this thesis indicate *LRRK2* rather than *MUC19* is likely to be responsible for the CD GWAS signal at the rs11175593 locus. In order to definitely prove this, further genotyping in the Edinburgh cohort concentrating on SNPs at the 5' end of the *LRRK2* gene would be necessary. Should this prove an association with IBD, genotyping across the exons would be a further useful step in elucidating potentially 'causative' mutations.

If a functioning LRRK2 antibody was available, IHC and immunofluorescence could be used to examine LRRK2 protein expression and localization in gastrointestinal biopsy samples, both with and without inflammation. Quantitative PCR on patient samples would also be useful: given the potential role of LRRK2 in the IF-γ mediated immune response[322], PBMCs isolated from CD patients and controls should be used examined for *LRRK2* mRNA expression as well as colonic tissues samples.

Chapter 10 Appendices

10.1 Invitation letter – Dundee IBD recruitment

Dear

Research project into genetic factors in inflammatory bowel disease-
Can you help?

We would like to invite you to take part in our research project, which is aimed at identifying genetic factors that influence the severity of inflammatory bowel disease. It is well known that the outcome for patients with ulcerative colitis/Crohn's disease is very variable - some patients have very mild disease with few if any symptoms throughout their life. A proportion, however, develop severe symptoms, and require intensive hospital treatment, or even surgery. At the present time there is a real need to identify factors which might predict the natural history of disease, and help identify those patients who are likely to develop severe symptoms.

In recent years progress has been made in identifying genetic markers of severity of ulcerative colitis and Crohn's disease.

The study is outlined in the attached information sheet. We ask that you would be prepared to give a blood or saliva sample for genetic analysis. A Doctor or Nurse, who would also read carefully through your hospital notes, would take the blood. The blood/saliva sample would be used for studies of genetic markers, which may help predict the course of ulcerative colitis/Crohn's disease.

All data and DNA samples would be stored securely, and the data and samples would be coded and therefore anonymous.

I hope that you are able to help with this research, which is promising to provide progress in understanding the cause of inflammatory bowel disease, and may lead to an improvement in treatment of disease.

If you are prepared to take part, please contact our research nurse
Shirley McLeod:
tel: 01382 660111, extension 35302 or

email: shirley.mcleod@nhs.net or

approach Shirley directly at the out-patient clinic

We could arrange to see you at a convenient time. Alternatively, we could see you at your next clinic appointment on________________.

Thank you very much for considering to help.

With best wishes.

Yours sincerely

Craig Mowat Consultant Gastroenterologist	Nigel Reynolds Consultant Gastroenterologist

10.2 Patient information sheet - Dundee IBD recruitment

Patient / Volunteer Information Sheet
AN INVESTIGATION INTO THE GENETIC DETERMINANTS OF SUSCEPTIBILITY AND PROGRESSION IN CHRONIC INFLAMMATORY BOWEL DISEASE.

We would like to invite you to take part in a research project. We believe it to be of potential importance. However, before you decide whether or not you wish to participate, we need to be sure that you understand firstly why we are doing it, and secondly what it would involve if you agreed. We are therefore providing you with the following information. Read it carefully and be sure to ask any questions you have, and, if you want, discuss it with outsiders. We will do our best to explain and to provide any further information you may ask for now or later. You do not have to make an immediate decision.

Background to the study
The chronic inflammatory bowel diseases, Crohn's disease and ulcerative colitis are common causes of gastrointestinal illness in Western Europe. They are particularly common in the North East of Scotland. The diseases cause considerable illness, particularly affecting quality of life in young people. Genetic studies, examining twins and siblings, have provided strong evidence that both genetic and environmental factors are important in determining who develops the disease (susceptibility). Genetic factors appear to be stronger in Crohn's disease than in ulcerative colitis. In recent years considerable progress has been made in identifying regions of the human genome which are associated with susceptibility to inflammatory bowel disease and are therefore likely to contain true genes involved in determining disease susceptibility. The regions of the genome that have been identified to date are large, and now need to be narrowed down before the genes may be identified. A number of strategies are currently being applied in attempts to narrow these regions. The establishment of a DNA bank, created by taking blood from affected patients and healthy controls will allow further study of genetic markers with the overall objective of identifying susceptibility genes.
The overall aims and objectives of this project are to identify genes involved in determining susceptibility and disease progression in the chronic inflammatory bowel diseases, Crohn's disease and ulcerative colitis. We aim to do this by comparing the DNA of patients with ulcerative colitis/Crohn's disease against the DNA of healthy volunteers, and by comparing the DNA of patients with mild forms

of ulcerative colitis/Crohn's disease with the DNA of those patients with more severe disease.
We hope that the results of this research will:
increase our understanding of the cause of Crohn's disease and ulcerative colitis.
provide a better understanding of the natural course of Crohn's disease and ulcerative colitis.
Allow better use of medical and surgical therapy.

The Tayside Committee on Medical Research Ethics, which has responsibility for scrutinising all proposals for medical research on humans in Tayside, has examined the proposal and has raised no objections from the point of view of medical ethics. If you agree to participate, monitors from the Committee may wish to examine your research records.

You have been chosen as a possible participant because: (tick box)

☐ You have Crohn's disease
☐ You have ulcerative colitis
☐ A patient with ulcerative colitis/Crohn's disease has nominated you as a healthy volunteer

What are you asking me to do?
To help with the research we would ask you to:
Complete a brief questionnaire
Donate a small blood sample (4 teaspoons, 20ml of blood) or saliva sample so that we can extract genetic material (DNA) from the blood. The DNA would be used to search for genetic changes that might predict the development or progress of ulcerative colitis/Crohn's disease. We would arrange for the blood to taken at a convenient time. This carries no significant risk, although there may be a very small bruise.
Sign a consent form for the study.

What will happen to my blood sample?
The blood/saliva sample will be given a code number so that laboratory personnel will not know your identity.
The details of the genetic analysis will only be known to medically qualified investigators involved in the study and will not be available to anyone else.
All data will be stored in secure databases.
The research does not constitute "a genetic test" as defined by insurance companies.
The genetic make-up of a patient with ulcerative colitis/Crohn's disease is estimated to account for 20% of the underlying cause. Many genes are believed to be implicated and this area of research is at a very early stage.
If, at a later date, we wished to contact your relatives, we would only do this with your permission.
Your blood/saliva sample may be used again in the future if new tests become available to further investigate the genetic basis of ulcerative colitis/Crohn's disease.

Future studies investigating the genetic basis of ulcerative colitis/Crohn's disease may be performed in collaboration with commercial companies. Collaboration with commercial companies provides us with much needed resources required to develop our research and to develop new treatment for inflammatory bowel disease. Although such companies would have access to information provided by DNA samples, including your own, they will never be able to identify you or link you to such information.

What will happen to the other information collected in the study?
All information collected will be held on a secure database which can be accessed only by the researchers involved in the study.

When is the research likely to benefit patients?
There has been great progress in identifying genetic markers, and clinical factors, which predict the course of ulcerative colitis/Crohn's disease. We very much hope that within a few years, it will be possible to identify those patients who are likely to suffer severe forms, and adjust medical therapy accordingly.
Will my GP know that I am involved in this study?
Yes, we shall notify your GP about your participation.
Who should I talk to if I have any questions or concerns?
Please feel free to contact Dr Craig Mowat, Consultant Gastroenterologist, or Dr Nigel Reynolds, Consultant Gastroenterologist, Wards 5&6, Ninewells Hospital, Dundee DD1 9SY, Telephone Number: 01382 660111.

Can I choose not to take part?
Yes of course. Participation in this study is entirely voluntary and you are free to refuse to take part or to withdraw from the study at any time without having to give a reason and without this affecting your future medical care or your relationship with medical staff looking after you.

10.3 Patient Questionnaire - Dundee IBD recruitment

Patient Identification Number

(for lab use only)

Genetics of Inflammatory Bowel Disease

Confidential

QUESTIONNAIRE FOR PATIENTS WITH CROHN'S DISEASE OR ULCERATIVE COLITIS

Please fill in this questionnaire, and bring it with you to your next clinic visit. Our Research Nurse in the clinic will answer any questions which you may have, and will help you fill it in if necessary.

Sex: M/F
(Please delete as appropriate)

Date of Birth:

Diagnosis: (please delete as appropriate)
Ulcerative Colitis/Crohn's disease

Date of diagnosis (year):

Name of present hospital consultant

Other medical problems (please list):

Have you ever had any trouble with your joints? (delete as appropriate) **YES/NO**

Have you ever been admitted to hospital with a flare up of your ulcerative colitis?
YES/NO *(for ulcerative colitis patients only)*

Smoking

Do you smoke cigarettes? (delete as appropriate) YES/NO

If Yes,

How many cigarettes a day?

When did you start smoking?

If No,

Have you ever smoked? (delete as appropriate) YES/NO

When did you start smoking? (Year)

When did you stop? (Year)

How many cigarettes a day?

Family History

1. In total, how many brothers sisters do you have?

2. How many children do you and have? (put 0 if no children)

3. How many living first degree relatives (parents, siblings, children) do you have?

3. Do any family members (grandparents, parents, aunts, uncles, siblings or children) have:

Crohn’s disease? **YES/NO**

Ulcerative colitis? **YES/NO**

Coeliac disease? **YES/NO**

Colon cancer? **YES/NO**

Please give details of affected relatives in the space provided including their age at onset of disease if known:

4. Ethnic origin

Which of the following best describes your ethnic origin?

White European
Jewish
Japanese
Other/Unknown
Hispanic
Afro-Caribbean
Asian

Surgery

Have you had any operations due to your Crohn's disease or ulcerative colitis?

If yes, please list below, with year if possible.

Have you had your tonsils removed? If yes, when? (year)

Have you had your appendix removed? If yes, when? (year)

Medications

Have you ever been treated with Azathioprine? **YES / NO / Don't know**

If yes, were you tolerant or intolerant of this medication?
(please tick as appropriate)

Are you taking Azathioprine at present? **YES / NO**

Have you ever been treated with Infliximab? **YES / NO**

Oral contraceptive status **Current Never Ex-**
(please circle)

10.4 Consent form - Dundee IBD recruitment

NHS Tayside

An investigation into the genetic determinants of susceptibility and progression in chronic inflammatory bowel disease

CONSENT FORM

Have you read and understood the Participant Information Sheet? Yes ☐ No ☐

Have you been given the opportunity to ask questions & further discuss the study? Yes ☐ No ☐

Have you received satisfactory answers to all of your questions? Yes ☐ No ☐

Have you now received enough information about the study? Yes ☐ No ☐

Do you understand that your participation is entirely voluntary? Yes ☐ No ☐

Do you understand that you are free to withdraw from this study:

At any time? Yes ☐ No ☐

Without having to give a reason for withdrawing? Yes ☐ No ☐

Without this affecting your present or future care? Yes ☐ No ☐

Do you agree to take part in this study? Yes ☐ No ☐

Do you agree to the DNA being used for future research? Yes ☐ No ☐

Note that it is a requirement that your research records and, if necessary, your medical records are available for scrutiny by monitors of the sponsor organisation and that by signing this document you are giving your consent to this.

Participant's signature________________________________Date_________________

Participant's name in block capital letters_____________________________________

Telephone contact (participant) Home___________________Work________________

Signature witnessed by_______________________________Date_________________

Witness name in block capital letters____

10.5 Crohn's Disease Clinical Data Form

Demographics

Lab Number

(In absence of sticky label please record
full name, sex, date of birth and

Physician

Ethnicity

Date of Onset of Symptoms

Date of Diagnosis

Oral contraceptive status
Current Never Ex (please circle)

<u>Smoking status</u>
Current Never Ex (please circle)

<u>Smoking at diagnosis</u> yes no ex don't know

Smoking amount
0-4 5-14 15-24 25+ (please circle)

Smoking at follow-up yes no ex don't know
Number of cigarettes per day =

Family history of IBD
Yes No (please circle)

Details of affected relatives, their diagnosis and age of onset.

Form completed by

Signature …………………………………… Date ……………………

Location please tick all that apply below

At diagnosis

Investigations UGIE BaFT Ba enema Colonoscopy Sigmoidoscopy

Oral OGD Jejunal Ileal TI Colonic Rectal Anal/Perianal

Maximal extent prior to 1st resection Montreal classification (please circle)

L1 L2 L3 L4 L1+L4 L2+L4 L3+L4

<u>At latest follow up</u> Date Investigation

Oral OGD Jejunal Ileal TI Colonic Rectal Anal/Perianal

Stoma Ileal Colonic

<u>Behaviour</u> <u>Montreal classification</u> (please circle)

<u>At diagnosis</u> Inflammatory B1 Stricturing B2 Penetrating B3

Or any of these plus penetrating perianal disease B1p B2p B3p

If stricture/fistula, give site + evidence (? symptomatic, investigation, treatment)

Date of 1st behaviour change
If stricture/fistula, give site, evidence (? symptomatic, investigation, treatment)

Date of 2nd behaviour change
Details

<u>At 5 years of F-UP</u> B1 B2 B3 B1p B2p B3p

<u>At 10 years of F-UP</u> B1 B2 B3 B1p B2p B3p

<u>At latest follow up</u> B1 B2 B3 B1p B2p B3p

Behaviour interim? Yes / No

Extra-intestinal manifestations please tick

Joints large joint related to disease activity
small joint unrelated to disease activity
AS
Sacro-ilitis

Skin Erythema nodosum
Pyoderma
Psoriasis
Mouth ulcers

Eyes Uveitis
Episcleritis
Conjunctivitis
Undiagnosed ocular inflammation

Liver PSC

Cancer Colon
Other (specify)

Other illnesses (specify)

<u>Surgical Details:</u> List date and operation details

1st operation Date

2nd operation Date

3rd operation Date

4th operation Date

5th operation Date

6th operation Date

Crohn's Disease Behaviour Data Form

DATE OF DIAGNOSIS:

Disease Location

Date	Oral	OgD	Jej	Ileal	TI	Colonic	Rect	Anal	Notes?

Disease Behaviour

Date	Inflam	Strict	Penetrate	Notes?

Date of Last Clinic: Completed By:

10.6 Ulcerative Colitis Clinical Data Form

Demographics | Lab Number

Diagnosis: UC Indeterminate colitis

Physician:

Ethnicity:

Date of Onset of Symptoms:

Date of Diagnosis:

Oral contraceptive status
Current Never Ex (please circle)

<u>Smoking status</u> **<u>Smoking at diagnosis</u>** yes no ex don't know
Current Never Ex (please circle)

<u>Smoking amount</u> **<u>Smoking at follow-up</u>** yes no ex don't know

0-4 5-14 15-24 25+ (please circle) Number of cigarettes per day =

Family history of IBD
Yes No (please circle)

Details of affected relatives, their diagnosis and age of onset.

Form completed by

Signature …………………………………… Date ……………………

Extent of disease at diagnosis: please circle

Rectum only	Colonoscopic	Histological	Radiological
Recto-sigmoid	Colonoscopic	Histological	Radiological
Splenic flexure	Colonoscopic	Histological	Radiological
Hepatic flexure	Colonoscopic	Histological	Radiological
Total	Colonoscopic	Histological	Radiological

Investigations undertaken:

If extent is unknown, please note why:

Montreal Extent at Diagnosis E1 E2 E3

Montreal Maximal Extent During Follow-up E1 E2 E3

Extent of disease at last follow-up investigation:

Date of Investigation: Type of Investigation:

Rectum only	Colonoscopic	Histological	Radiological
Recto-sigmoid	Colonoscopic	Histological	Radiological
Splenic flexure	Colonoscopic	Histological	Radiological
Hepatic flexure	Colonoscopic	Histological	Radiological
Total	Colonoscopic	Histological	Radiological

<u>Please note below</u>, if there have been no further investigations since diagnosis or if investigations undertaken were normal or whether the data is unavailable:

Montreal Extent at Follow-up E1 E2 E3

Colectomy Yes / No Reason: Severe disease

Chronic continuous

Date of Colectomy:

Dysplasia/cancer

Ileo-anal Pouch Yes / No Pouchitis Yes
- acute intermittent
- chronic

No

Extra-intestinal manifestations please tick

Joints large joint related to disease activity
small joint unrelated to disease activity
AS
Sacro-ilitis

Skin Erythema nodosum
Pyoderma
Psoriasis
Mouth ulcers

Eyes Uveitis
Episcleritis
Conjunctivitis
Undiagnosed ocular inflammation

Liver PSC

Cancer Colon

Other(specify)

Other illnesses (specify)

Last Clinic Date:

10.7 Patient biopsy information letter

GASTROINTESTINAL UNIT
Molecular Medicine Centre
The University of Edinburgh
Western General Hospital
Edinburgh EH4 2XU
Telephone – 0131 651 1807
Fax – 0131 651 1085
Email anne.phillips@ed.ac.uk

Dear

Re: Research into the genetics of inflammatory bowel disease

In the Gastrointestinal Unit we are researching the genetics of inflammatory bowel disease, and hope that you might be willing to take part in one of our studies.

I understand that you are having a colonoscopy soon. When you have your test, you will have a number of tiny samples of tissue (biopsies) taken from the lining of your bowel for your clinical care. The number taken will depend on the reason for doing the colonoscopy, but could be up to 30. Each of these samples is less than the size of a small grain of rice. What we would like to do, with your permission, is take an extra 5-10 samples for research purposes. Taking these extra samples would add less than 5 minutes to the procedure. The aim of our study is to help us further our understanding of the processes that occur in both Crohn's disease and ulcerative colitis, in the hope that in the future we might be able to develop more effective treatments or get better at using the ones we have.

As you will be aware, a colonoscopy is generally an extremely safe procedure, but is not risk-free. By taking a few more samples there could be a very slightly increased risk of minor problems, which would usually settle on their own without requiring any further action.

If you agree to take part, I would be most grateful if you could fill in the enclosed form and bring it with you on the day. If you choose not to participate then please rest assured that your medical care will not be affected in any way.

If you have any questions I would be happy to talk to you about the study - I can be e-mailed or phoned using the above details. Alternatively, Dr Ian Arnott, who is the independent medical adviser for this study, can be contacted on 0131 537 3115.

Yours sincerely

Dr Anne Phillips
Clinical Research Fellow

10.8 Control biopsy information letter

GASTROINTESTINAL UNIT
Molecular Medicine Centre
The University of Edinburgh
Western General Hospital
Edinburgh EH4 2XU
Telephone – 0131 651 1807
Fax – 0131 651 1085
Email anne.phillips@ed.ac.uk

Dear ,

Re: Research into the genetics of inflammatory bowel disease

In the Gastrointestinal Unit we are researching the genetics of inflammatory bowel disease (Crohn's disease and ulcerative colitis) and hope that you might be willing to take part in one of our studies. These diseases affect the gut and are a common cause of ill health in Scotland including weight loss, diarrhoea and the need for operations. The causes of Crohn's disease and ulcerative colitis are not yet fully understood.

I understand that you do not have inflammatory bowel disease and I would like to enter you into the control group for this study. Control groups are vital for these studies so that we can compare people with and without inflammatory bowel disease.

I am aware that you are having a colonoscopy soon. When you have your test, you are likely to have a number of tiny samples of tissue (biopsies) taken from the lining of the bowel for your clinical care. The number taken will depend on the reason for doing the colonoscopy, but could be up to 30. Each of these samples is less than the size of a small grain of rice. What we would like to do, with your permission, is take an extra 5-10 samples for research purposes. Taking these extra samples would add less than 5 minutes to the procedure. The aim of our study is to help us further our understanding of the processes that occur in both Crohn's disease and ulcerative colitis, in the hope that in the future we might be able to develop more effective treatments or get better at using the ones we have.

As you will be aware, a colonoscopy is generally an extremely safe procedure, but is not risk-free. By taking a few more samples there could be a very slightly increased risk of minor problems, which would usually settle on their own without requiring any further action.

If you agree to take part, I would be most grateful if you could fill in the enclosed form and bring it with you on the day. If you choose not to participate then please rest assured that your medical care will not be affected in any way.

If you have any questions I would be happy to talk to you about the study - I can be e-mailed or phoned using the above details. Alternatively, Dr Ian Arnott, who is the independent medical adviser for this study, can be contacted on 0131 537 3115.

Yours sincerely,

Dr Anne Phillips
Research Registrar

10.9 Biopsy consent form

Consent for colonoscopy: IBD patients

Investigation of the expression of immune response genes in the gastro-intestinal tract in inflammatory bowel disease.

I ___________________________ (please print your name in block capitals) have read and understood the patient information sheet.

I understand that my participation in the research is entirely voluntary and does not affect my future medical care.

I understand that that the research will involve:

- allowing analysis of biopsies taken at colonoscopy
- filling in a short questionnaire
- allowing trained medical personnel to read my medical case notes

Participant	Witness
Signed ……………………	Name …………………………
Date ………………………	Signed ………………….Date…….

Chapter 11 Bibliography

(1) Eaden JA, Abrams KR, Mayberry JF. The risk of colorectal cancer in ulcerative colitis: a meta-analysis. Gut 2001; 48(4):526-535.

(2) Lennard-Jones JE. Classification of inflammatory bowel disease. Scandinavian Journal of Gastroenterology 1989; 24(170 (Suppl.)):2-6.

(3) Shivananda S, Lennard-Jones J, Logan R, Fear N, Price A, Carpenter L et al. Incidence of inflammatory bowel disease across Europe: is there a difference between north and south? Results of the European Collaborative Study on Inflammatory Bowel Disease (EC-IBD). Gut 1996; 39(5):690-697.

(4) Fellows IW, Freeman JG, Holmes GK. Crohn's disease in the city of Derby, 1951-85. Gut 1990; 31(11):1262-1265.

(5) Loftus CG, Loftus EV, Jr., Harmsen WS, Zinsmeister AR, Tremaine WJ, Melton LJ, III et al. Update on the incidence and prevalence of Crohn's disease and ulcerative colitis in Olmsted County, Minnesota, 1940-2000. Inflamm Bowel Dis 2007; 13(3):254-261.

(6) Bernstein CN, Wajda A, Svenson LW, MacKenzie A, Koehoorn M, Jackson M et al. The epidemiology of inflammatory bowel disease in Canada: a population-based study. Am J Gastroenterol 2006; 101(7):1559-1568.

(7) Lindberg E, Jornerot G. The incidence of Crohn's disease is not decreasing in Sweden. Scand J Gastroenterol 1991; 26(5):495-500.

(8) Bodger K, Halfvarson J, Dodson AR, Campbell F, Wilson S, Lee R et al. Altered colonic glycoprotein expression in unaffected monozygotic twins of inflammatory bowel disease patients. Gut 2006; 55(7):973-977.

(9) Gower-Rousseau C, Salomez JL, Dupas JL, Marti R, Nuttens MC, Votte A et al. Incidence of inflammatory bowel disease in northern France (1988-1990). Gut 1994; 35(10):1433-1438.

(10) Lapidus A. Crohn's disease in Stockholm County during 1990-2001: an epidemiological update. World Journal of Gastroenterology 2006; 12(1):75-81.

(11) Lashner BA, Loftus EV, Jr. True or false? The hygiene hypothesis for Crohn's disease. American Journal of Gastroenterology 2006; 101(5):1003-1004.

(12) Odes S, Vardi H, Friger M, Wolters F, Russel MG, Riis L et al. Cost analysis and cost determinants in a European inflammatory bowel disease inception cohort with 10 years of follow-up evaluation. Gastroenterology 2006; 131(3):719-728.

(13) Steed H, Walsh S, Reynolds N. Crohn's disease incidence in NHS Tayside. Scott Med J 2010; 55(3):22-25.

(14) Farmer RG, Hawk WA, Turnbull RB, Jr. Clinical patterns in Crohn's disease: a statistical study of 615 cases. Gastroenterology 1975; 68(4 Pt 1):627-635.

(15) Greenstein AJ, Lachman P, Sachar DB, Springhorn J, Heimann T, Janowitz HD et al. Perforating and non-perforating indications for repeated operations in Crohn's disease: evidence for two clinical forms. Gut 1988; 29(5):588-592.

(16) Sachar DB, Andrews HA, Farmer RG, Pallone F, Pena AS, Prantera C et al. Proposed classification of patient subgroups in Crohn's disease. Gastroenterol Intl 1992; 5:141-154.

(17) Gasche C, Scholmerich J, Brynskov J, D'Haens G, Hanauer SB, Irvine EJ et al. A simple classification of Crohn's disease: report of the Working Party for the World Congresses of Gastroenterology, Vienna 1998. Inflamm Bowel Dis 2000; 6(1):8-15.

(18) Sachar DB, Bodian CA, Goldstein ES, Present DH, Bayless TM, Picco M et al. Is perianal Crohn's disease associated with intestinal fistulization? Am J Gastroenterol 2005; 100(7):1547-1549.

(19) Smith BRKB, Arnott IDRM, Drummond HEB, Nimmo ERP, Satsangi JD. Disease Location, Anti-Saccharomyces cerevisiae Antibody, and NOD2/CARD15 Genotype Influence the Progression of Disease Behavior in Crohn's Disease. Inflammatory Bowel Diseases 2004; 10(5):521-528.

(20) Silverberg MS, Satsangi J, Ahmad T, Arnott ID, Bernstein CN, Brant SR et al. Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: Report of a Working Party of the 2005 Montreal World Congress of Gastroenterology. Can J Gastroenterol 2005; 19 Suppl A:5-36.

(21) Levine A, Griffiths A, Markowitz J, Wilson DC, Turner D, Russell RK et al. Pediatric modification of the Montreal classification for inflammatory bowel disease: The Paris classification. Inflamm Bowel Dis 2010.

(22) Truelove SC, WITTS LJ. Cortisone in ulcerative colitis; preliminary report on a therapeutic trial. Br Med J 1954; 2(4884):375-378.

(23) Truelove SC, WITTS LJ. Cortisone in ulcerative colitis; final report on a therapeutic trial. Br Med J 1955; 2(4947):1041-1048.

(24) Desreumaux P, Ghosh S. Review article: mode of action and delivery of 5-aminosalicylic acid - new evidence. Aliment Pharmacol Ther 2006; 24 Suppl 1:2-9.

(25) Sutherland L, Macdonald JK. Oral 5-aminosalicylic acid for maintenance of remission in ulcerative colitis. Cochrane Database Syst Rev 2006;(2):CD000544.

(26) Eaden J. Review article: the data supporting a role for aminosalicylates in the chemoprevention of colorectal cancer in patients with inflammatory bowel disease. Aliment Pharmacol Ther 2003; 18 Suppl 2:15-21.

(27) Sutherland L, Macdonald JK. Oral 5-aminosalicylic acid for induction of remission in ulcerative colitis. Cochrane Database Syst Rev 2006;(2):CD000543.

(28) Marshall JK, Thabane M, Steinhart AH, Newman JR, Anand A, Irvine EJ. Rectal 5-aminosalicylic acid for induction of remission in ulcerative colitis. Cochrane Database Syst Rev 2010;(1):CD004115.

(29) Akobeng AK, Gardener E. Oral 5-aminosalicylic acid for maintenance of medically-induced remission in Crohn's Disease. Cochrane Database Syst Rev 2005;(1):CD003715.

(30) Sahasranaman S, Howard D, Roy S. Clinical pharmacology and pharmacogenetics of thiopurines. Eur J Clin Pharmacol 2008; 64(8):753-767.

(31) Prefontaine E, Macdonald JK, Sutherland LR. Azathioprine or 6-mercaptopurine for induction of remission in Crohn's disease. Cochrane Database Syst Rev 2010;(6):CD000545.

(32) Timmer A, McDonald JW, Macdonald JK. Azathioprine and 6-mercaptopurine for maintenance of remission in ulcerative colitis. Cochrane Database Syst Rev 2007;(1):CD000478.

(33) Gisbert JP, Linares PM, McNicholl AG, Mate J, Gomollon F. Meta-analysis: the efficacy of azathioprine and mercaptopurine in ulcerative colitis. Aliment Pharmacol Ther 2009; 30(2):126-137.

(34) Komatsu M, Kobayashi D, Saito K, Furuya D, Yagihashi A, Araake H et al. Tumor necrosis factor-alpha in serum of patients with inflammatory bowel disease as measured by a highly sensitive immuno-PCR. Clin Chem 2001; 47(7):1297-1301.

(35) D'Haens G, van Deventer S, Van Hogezand R, Chalmers D, Kothe C, Baert F et al. Endoscopic and histological healing with infliximab anti-tumor necrosis factor antibodies in Crohn's disease: A European multicenter trial. Gastroenterology 1999; 116(5):1029-1034.

(36) Present DH, Rutgeerts P, Targan S, Hanauer SB, Mayer L, van Hogezand RA et al. Infliximab for the treatment of fistulas in patients with Crohn's disease. N Engl J Med 1999; 340(18):1398-1405.

(37) Targan SR, Hanauer SB, van Deventer SJ, Mayer L, Present DH, Braakman T et al. A short-term study of chimeric monoclonal antibody cA2 to tumor necrosis factor alpha for Crohn's disease. Crohn's Disease cA2 Study Group. N Engl J Med 1997; 337(15):1029-1035.

(38) Hanauer SB, Feagan BG, Lichtenstein GR, Mayer LF, Schreiber S, Colombel JF et al. Maintenance infliximab for Crohn's disease: the ACCENT I randomised trial. Lancet 2002; 359(9317):1541-1549.

(39) Rutgeerts P, Sandborn WJ, Feagan BG, Reinisch W, Olson A, Johanns J et al. Infliximab for induction and maintenance therapy for ulcerative colitis. N Engl J Med 2005; 353(23):2462-2476.

(40) Baert F, Noman M, Vermeire S, Van AG, D' HG, Carbonez A et al. Influence of immunogenicity on the long-term efficacy of infliximab in Crohn's disease. N Engl J Med 2003; 348(7):601-608.

(41) Hanauer SB, Sandborn WJ, Rutgeerts P, Fedorak RN, Lukas M, MacIntosh D et al. Human anti-tumor necrosis factor monoclonal antibody (adalimumab) in Crohn's disease: the CLASSIC-I trial. Gastroenterology 2006; 130(2):323-333.

(42) Sandborn WJ, Hanauer SB, Rutgeerts P, Fedorak RN, Lukas M, MacIntosh DG et al. Adalimumab for maintenance treatment of Crohn's disease: results of the CLASSIC II trial. Gut 2007; 56(9):1232-1239.

(43) Colombel JF, Sandborn WJ, Rutgeerts P, Enns R, Hanauer SB, Panaccione R et al. Adalimumab for maintenance of clinical response and remission in patients with Crohn's disease: the CHARM trial. Gastroenterology 2007; 132(1):52-65.

(44) Sandborn WJ, Rutgeerts P, Enns R, Hanauer SB, Colombel JF, Panaccione R et al. Adalimumab induction therapy for Crohn disease previously treated with infliximab: a randomized trial. Ann Intern Med 2007; 146(12):829-838.

(45) Lichtenstein GR, Feagan BG, Cohen RD, Salzberg BA, Diamond RH, Chen DM et al. Serious infections and mortality in association with therapies for Crohn's disease: TREAT registry. Clin Gastroenterol Hepatol 2006; 4(5):621-630.

(46) Thayu M, Markowitz JE, Mamula P, Russo PA, Muinos WI, Baldassano RN. Hepatosplenic T-cell lymphoma in an adolescent patient after immunomodulator and biologic therapy for Crohn disease. Journal of Pediatric Gastroenterology & Nutrition 2005; 40(2):220-222.

(47) Melichar B, Bures J, Dedic K. Anorectal carcinoma after infliximab therapy in Crohn's disease: report of a case. Dis Colon Rectum 2006; 49(8):1228-1233.

(48) Rennard SI, Fogarty C, Kelsen S, Long W, Ramsdell J, Allison J et al. The safety and efficacy of infliximab in moderate to severe chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2007; 175(9):926-934.

(49) Lees CW, Ironside J, Wallace WA, Satsangi J. Resolution of non-small-cell lung cancer after withdrawal of anti-TNF therapy. N Engl J Med 2008; 359(3):320-321.

(50) Colombel JF, Sandborn WJ, Reinisch W, Mantzaris GJ, Kornbluth A, Rachmilewitz D et al. Infliximab, azathioprine, or combination therapy for Crohn's disease. N Engl J Med 2010; 362(15):1383-1395.

(51) Yacyshyn B, Chey WY, Wedel MK, Yu RZ, Paul D, Chuang E. A randomized, double-masked, placebo-controlled study of alicaforsen, an antisense inhibitor of intercellular adhesion molecule 1, for the treatment of subjects with active Crohn's disease. Clin Gastroenterol Hepatol 2007; 5(2):215-220.

(52) Mannon PJ, Fuss IJ, Mayer L, Elson CO, Sandborn WJ, Present D et al. Anti-interleukin-12 antibody for active Crohn's disease. New England Journal of Medicine 2004; 351(20):2069-2079.

(53) Sandborn WJ, Feagan BG, Fedorak RN, Scherl E, Fleisher MR, Katz S et al. A randomized trial of Ustekinumab, a human interleukin-12/23 monoclonal antibody, in patients with moderate-to-severe Crohn's disease. Gastroenterology 2008; 135(4):1130-1141.

(54) Oyama Y, Craig RM, Traynor AE, Quigley K, Statkute L, Halverson A et al. Autologous hematopoietic stem cell transplantation in patients with refractory Crohn's disease. Gastroenterology 2005; 128(3):552-563.

(55) Leijonmarck CE, Persson PG, Hellers G. Factors affecting colectomy rate in ulcerative colitis: an epidemiologic study. Gut 1990; 31(3):329-333.

(56) Ho GT, Mowat C, Goddard CJ, Fennell JM, Shah NB, Prescott RJ et al. Predicting the outcome of severe ulcerative colitis: development of a novel risk score to aid early selection of patients for second-line medical therapy or surgery. Aliment Pharmacol Ther 2004; 19(10):1079-1087.

(57) Travis SP, Farrant JM, Ricketts C, Nolan DJ, Mortensen NM, Kettlewell MG et al. Predicting outcome in severe ulcerative colitis. Gut 1996; 38(6):905-910.

(58) Mayberry JF, Rhodes J, Newcombe RG. Familial prevalence of inflammatory bowel disease in relatives of patients with Crohn's disease. Br Med J 1980; 280(6207):84.

(59) Halfvarson J, Bodin L, Tysk C, Lindberg E, Jarnerot G. Inflammatory bowel disease in a Swedish twin cohort: A long-term follow-up of concordance and clinical characteristics. Gastroenterology 2003; 124(7):1767-1773.

(60) Thompson NP, Driscoll R, Pounder RE, Wakefield AJ. Genetics versus environment in inflammatory bowel disease: Results of a British twin study. British Medical Journal 1996; 312(7023):95-96.

(61) Orholm M, Binder V, Sorensen TIA, Rasmussen LP, Kyvik KO. Concordance of inflammatory bowel disease among Danish twins - Results of a nationwide study. Scandinavian Journal of Gastroenterology 2000; 35(10):1075-1081.

(62) Hampe J, Schreiber S, Shaw SH, Lau KF, Bridger S, MacPherson AJ et al. A genomewide analysis provides evidence for novel linkages in inflammatory bowel disease in a large European cohort. American Journal of Human Genetics 1999; 64(3):808-816.

(63) Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, McLeod RS, Griffiths AM et al. Genomewide search in Canadian families with inflammatory bowel disease reveals two novel susceptibility loci. American Journal of Human Genetics 2000; 66(6):1863-1870.

(64) Ma Y, Ohmen JD, Li Z, Bentley LG, McElree C, Pressman S et al. A genome-wide search identifies potential new susceptibility loci for Crohn's disease. Inflamm Bowel Dis 1999; 5(4):271-278.

(65) Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 2001; 411(6837):599-603.

(66) Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R et al. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 2001; 411(6837):603-606.

(67) Kruglyak L. The road to genome-wide association studies. Nature Reviews Genetics 2008; 9(4):314-318.

(68) Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ et al. A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. Science 2006; 314(5804):1461-1463.

(69) Imielinski M, Baldassano RN, Griffiths A, Russell RK, Annese V, Dubinsky M et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. Nat Genet 2009; 41(12):1335-1340.

(70) Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D et al. A novel susceptibility locus for Crohns disease identified by whole genome association maps to a gene desert on chromosome 5p13.1 and modulates the level of expression of the prostaglandin receptor EP4. PloS Genetics 2007; 3(4):e58.

(71) Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 2007; 39(5):596-604.

(72) McGovern DP, Jones MR, Taylor KD, Marciante K, Yan X, Dubinsky M et al. Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Hum Mol Genet 2010; 19(17):3468-3476.

(73) the Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007; 447(7145):661-678.

(74) Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 2010.

(75) McGovern DP, Gardet A, Torkvist L, Goyette P, Essers J, Taylor KD et al. Genome-wide association identifies multiple ulcerative colitis susceptibility loci. Nat Genet 2010; 42(4):332-337.

(76) Silverberg MS, Cho JH, Rioux JD, McGovern DP, Wu J, Annese V et al. Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. Nat Genet 2009; 41(2):216-220.

(77) Franke A, Balschun T, Sina C, Ellinghaus D, Hasler R, Mayr G et al. Genome-wide association study for ulcerative colitis identifies risk loci at 7q22 and 22q13 (IL17REL). Nat Genet 2010; 42(4):292-294.

(78) Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA, Phillips A et al. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. Nat Genet 2009; 41(12):1330-1334.

(79) Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 2010; 464(7289):713-720.

(80) Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet 2011; 43(3):246-252.

(81) Hugot JP, Laurent-Puig P, Gower-Rousseau C, Olson JM, Lee JC, Beaugerie L et al. Mapping of a susceptibility locus for Crohn's disease on chromosome 16. Nature 1996; 379(6568):821-823.

(82) Girardin SE, Boneca IG, Viala J, Chamaillard M, Labigne A, Thomas G et al. Nod2 is a general sensor of peptidoglycan through muramyl dipeptide (MDP) detection. Journal of Biological Chemistry 2003; 278(11):8869-8872.

(83) Lesage S, Zouali H, Cezard JP, Colombel JF, Belaiche J, Almer S et al. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in

612 patients with inflammatory bowel disease. American Journal of Human Genetics 2002; 70(4):845-857.

(84) Ahmad T, Armuzzi A, Bunce M, Mulcahy-Hawes K, Marshall SE, Orchard TR et al. The molecular classification of the clinical manifestations of Crohn's disease. Gastroenterology 2002; 122(4):854-866.

(85) Abreu MT, Taylor KD, Lin YC, Hang T, Gaiennie J, Landers CJ et al. Mutations in NOD2 are associated with fibrostenosing disease in patients with Crohn's disease. Gastroenterology 2002; 123(3):679-688.

(86) Vermeire S, Cousineau J, Dufresne L, Pellerin N, Bernier ML, Bitton A et al. NOD2/CARD15 mutations in Crohn's disease in Quebec: Prevalence and association with clinical phenotypes. Gastroenterology 2002; 122(4):A295-A296.

(87) Yamazaki K, Takazoe M, Tanaka T, Kazumori T, Nakamura Y. Absence of mutation in the NOD2/CARD15 gene among 483 Japanese patients with Crohn's disease. Journal of Human Genetics 2002; 47(9):469-472.

(88) Arnott IDR, Nimmo ER, Drummond HE, Fennell J, Smith BRK, MacKinlay E et al. NOD2/CARD15, TLR4 and CD14 mutations in Scottish and Irish Crohn's disease patients: Evidence for genetic heterogeneity within Europe? Genes & Immunity 2004; 5(5):417-425.

(89) Inohara N, Ogura Y, Fontalba A, Gutierrez O, Pons F, Crespo J et al. Host recognition of bacterial muramyl dipeptide mediated through NOD2: Implications for Crohn's disease. Journal of Biological Chemistry 2003; 278(8):5509-5512.

(90) Barnich N, Aguirre JE, Reinecker HC, Xavier R, Podolsky DK. Membrane recruitment of NOD2 in intestinal epithelial cells is essential for nuclear factor-{kappa}B activation in muramyl dipeptide recognition. J Cell Biol 2005; 170(1):21-26.

(91) Lala S, Ogura Y, Osborne C, Hor SY, Bromfield A, Davies S et al. Crohn's disease and the NOD2 gene: A role for paneth cells. Gastroenterology 2003; 125(1):47-57.

(92) Ogura Y, Inohara N, Benito A, Chen FF, Yamaoka S, Nunez G. Nod2, a Nod1/Apaf-1 family member that is restricted to monocytes and activates NF-kappaB. J Biol Chem 2001; 276(7):4812-4818.

(93) Wehkamp J, Salzman NH, Porter E, Nuding S, Weichenthal M, Petras RE et al. Reduced Paneth cell alpha-defensins in ileal Crohn's disease.

Proceedings of the National Academy of Sciences of the United States of America 2005; 102(50):18129-18134.

(94) Kelsall B. Getting to the guts of NOD2. Nat Med 2005; 11(4):383-384.

(95) Lecine P, Esmiol S, Metais JY, Nicoletti C, Nourry C, McDonald C et al. The NOD2-RICK complex signals from the plasma membrane. J Biol Chem 2007; 282(20):15197-15207.

(96) Marks DJ, Harbord MW, MacAllister R, Rahman FZ, Young J, Al Lazikani B et al. Defective acute inflammation in Crohn's disease: a clinical investigation. Lancet 2006; 367(9511):668-678.

(97) Sewell GW, Marks DJ, Segal AW. The immunopathogenesis of Crohn's disease: a three-stage model. Curr Opin Immunol 2009; 21(5):506-513.

(98) Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet 2007; 39(2):207-211.

(99) Cheng JF, Ning YJ, Zhang W, Lu ZH, Lin L. T300A polymorphism of ATG16L1 and susceptibility to inflammatory bowel diseases: a meta-analysis. World J Gastroenterol 2010; 16(10):1258-1266.

(100) Kuballa P, Huett A, Rioux JD, Daly MJ, Xavier RJ. Impaired autophagy of an intracellular pathogen induced by a Crohn's disease associated ATG16L1 variant. PLoS One 2008; 3(10):e3391.

(101) Singh SB, Davis AS, Taylor GA, Deretic V. Human IRGM induces autophagy to eliminate intracellular mycobacteria. Science 2006; 313:1438-1441.

(102) Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, Fisher SA et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet 2007; 39(7):830-832.

(103) McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet 2008; 40(9):1107-1112.

(104) Nakagawa I. Autophagy defends cells against invading group A Streptococcus. Science 2004; 306:1037-1040.

(105) Gutierrez MG, Master SS, Singh SB, Taylor GA, Colombo MI, Deretic V. Autophagy Is a Defense Mechanism Inhibiting BCG and Mycobacterium tuberculosis Survival in Infected Macrophages. Cell 2004; 119(6):753-766.

(106) Heath RJ, Xavier RJ. Autophagy, immunity and human disease. Curr Opin Gastroenterol 2009; 25(6):512-520.

(107) Xu Y, Jagannath C, Liu XD, Sharafkhaneh A, Kolodziejska KE, Eissa NT. Toll-like receptor 4 is a sensor for autophagy associated with innate immunity. Immunity 2007; 27(1):135-144.

(108) Delgado MA, Elmaoued RA, Davis AS, Kyei G, Deretic V. Toll-like receptors control autophagy. EMBO J 2008; 27(7):1110-1121.

(109) Travassos LH, Carneiro LA, Ramjeet M, Hussey S, Kim YG, Magalhaes JG et al. Nod1 and Nod2 direct autophagy by recruiting ATG16L1 to the plasma membrane at the site of bacterial entry. Nat Immunol 2010; 11(1):55-62.

(110) Harrington LE, Hatton RD, Mangan PR, Turner H, Murphy TL, Murphy KM et al. Interleukin 17-producing CD4+ effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages. Nature Immunology 2005; 6(11):1123-1132.

(111) Wiekowski MT, Leach MW, Evans EW, Sullivan L, Chen SC, Vassileva G et al. Ubiquitous transgenic expression of the IL-23 subunit p19 induces multiorgan inflammation, runting, infertility, and premature death. J Immunol 2001; 166(12):7563-7570.

(112) Ghilardi N, Kljavin N, Chen Q, Lucas S, Gurney AL, de Sauvage FJ. Compromised humoral and delayed-type hypersensitivity responses in IL-23-deficient mice. J Immunol 2004; 172(5):2827-2833.

(113) Ouyang W, Kolls JK, Zheng Y. The biological functions of T helper 17 cell effector cytokines in inflammation. Immunity 2008; 28(4):454-467.

(114) Schmechel S, Konrad A, Diegelmann J, Glas J, Wetzke M, Paschos E et al. Linking genetic susceptibility to Crohn's disease with Th17 cell function: IL-22 serum levels are increased in Crohn's disease and correlate with disease activity and IL23R genotype status. Inflamm Bowel Dis 2008; 14(2):204-212.

(115) Brand S. Crohn's disease: Th1, Th17 or both? The change of a paradigm: new immunological and genetic insights implicate Th17 cells in the pathogenesis of Crohn's disease. Gut 2009; 58(8):1152-1167.

(116) Fisher SA, Tremelling M, Anderson CA, Gwilliam R, Bumpstead S, Prescott NJ et al. Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease. Nat Genet 2008; 40(6):710-712.

(117) Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nature Genetics 2008; 40(8):955-962.

(118) Parham C, Chirica M, Timans J, Vaisberg E, Travis M, Cheung J et al. A receptor for the heterodimeric cytokine IL-23 is composed of IL-12Rbeta1 and a novel cytokine receptor subunit, IL-23R. J Immunol 2002; 168(11):5699-5708.

(119) Mehra NK, Kaur G. MHC-based vaccination approaches: progress and perspectives. Expert Rev Mol Med 2003; 5(7):1-17.

(120) Satsangi J, Welsh KI, Bunce M, Julier C, Farrant JM, Bell JI et al. Contribution of genes of the major histocompatibility complex to susceptibility and disease phenotype in inflammatory bowel disease. Lancet 1996; 347(9010):1212-1217.

(121) Stokkers PCF, Reitsma PH, Tytgat GNJ, Van Deventer SJH. HLA-DR and -DQ phenotypes in inflammatory bowel disease: a meta-analysis. Gut 1999; 45(3):395-401.

(122) Kugathasan S, Baldassano RN, Bradfield JP, Sleiman PM, Imielinski M, Guthery SL et al. Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. Nat Genet 2008; 40(10):1211-1215.

(123) Yacyshyn B, Maksymowych W, Bowen-Yacyshyn MB. Differences in P-glycoprotein-170 expression and activity between Crohn's disease and ulcerative colitis. Human Immunology 1999; 60(8):677-687.

(124) Panwala CM, Jones JC, Viney JL. A novel model of inflammatory bowel disease: mice deficient for the multiple drug resistance gene, mdr1a, spontaneously develop colitis. Journal of Immunology 1998; 161(10):5733-5744.

(125) Thiebaut F, Hanauske AR, Vonhoff DD. Evidence for Binding of Extrachromosomal Dna-Sequences to Nuclear Matrix Proteins in Multidrug-Resistant Kb-V1 Cells. FEBS Letters 1993; 319(1-2):133-137.

(126) Coon JS, Wang YZ, Bines SD, Markham PN, Chong ASF, Gebel HM. Multidrug Resistance Activity in Human-Lymphocytes. Human Immunology 1991; 32(2):134-140.

(127) Hoffmeyer S, Burk O, von Richter O, Arnold HP, Brockmoller J, Johne A et al. Functional polymorphisms of the human multidrug-resistance gene: Multiple sequence variations and correlation of one allele with P-glycoprotein expression and activity in vivo. Proceedings of the National Academy of Sciences of the United States of America 2000; 97(7):3473-3478.

(128) Schwab M, Schaeffeler E, Marx C, Fromm MF, Kaskas B, Metzler J et al. Association between the C3435T MDR1 gene polymorphism and susceptibility for ulcerative colitis. Gastroenterology 2003; 124(1):26-33.

(129) Potocnik U, Ferkolj I, Glavac D, Dean M. Polymorphisms in multidrug resistance 1 (MDR1) gene are associated with refractory Crohn disease and ulcerative colitis. Genes & Immunity 2004; 5(7):530-539.

(130) Ho GT, Nimmo ER, Tenesa A, Fennell J, Drummond H, Mowat C et al. Allelic variations of the multidrug resistance gene determine susceptibility and disease behavior in ulcerative colitis. Gastroenterology 2005; 128(2):288-296.

(131) Brant SR, Panhuysen CI, Nicolae D, Reddy DM, Bonen DK, Karaliukas R et al. MDR1 Ala893 polymorphism is associated with inflammatory bowel disease. American Journal of Human Genetics 2003; 73(6):1282-1292.

(132) Croucher PJ, Mascheretti S, Foelsch UR, Hampe J, Schreiber S. Lack of association between the C3435T MDR1 gene polymorphism and inflammatory bowel disease in two independent Northern European populations. Gastroenterology 2003; 125(6):1919-1920.

(133) Palmieri O, Latiano A, Valvano R, D'Inca R, Vecchi M, Sturniolo GC et al. Multidrug resistance 1 gene polymorphisms are not associated with inflammatory bowel disease and response to therapy in Italian patients. Alimentary Pharmacology & Therapeutics 2005; 22(11-12):1129-1138.

(134) Onnie CM, Fisher SA, Pattni R, Sanderson J, Forbes A, Lewis CM et al. Associations of allelic variants of the multidrug resistance gene (ABCB1 or MDR1) and inflammatory bowel disease and their effects on disease behavior: a case-control and meta-analysis study. Inflammatory Bowel Diseases 2006; 12(4):263-271.

(135) Ho GT, Soranzo N, Nimmo ER, Tenesa A, Goldstein DB, Satsangi J. ABCB1/MDR1 gene determines susceptibility and phenotype in ulcerative colitis: discrimination of critical variants using a gene-wide haplotype tagging approach. Hum Mol Genet 2006; 15(5):797-805.

(136) Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV et al. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science 2007; 315(5811):525-528.

(137) Nakamura H, Sudo T, Tsuiki H, Miyake H, Morisaki T, Sasaki J et al. Identification of a novel human homolog of the Drosophila dlg, P-dlg, specifically expressed in the gland tissues and interacting with p55. FEBS Letters 1998; 433(1-2):63-67.

(138) Stoll M, Corneliussen B, Costello CM, Waetzig GH, Mellgard B, Koch WA et al. Genetic variation in DLG5 is associated with inflammatory bowel disease. Nature Genetics 2004; 36(5):476-480.

(139) Daly MJ, Pearce AV, Farwell L, Fisher SA, Latiano A, Prescott NJ et al. Association of DLG5 R30Q variant with inflammatory bowel disease. European Journal of Human Genetics 2005; 13(7):835-839.

(140) Russell RK, Drummond HE, Nimmo ER, Anderson N, Wilson DC, Gillett PM et al. The contribution of the DLG5 113A variant in early-onset inflammatory bowel disease. Journal of Pediatrics 2007; 150(3):268-273.

(141) Tenesa A, Noble C, Satsangi J, Dunlop M. Association of DLG5 and inflammatory bowel disease across populations. European Journal of Human Genetics 2006; 14(3):259-260.

(142) Torok HP, Glas J, Tonenchi L, Lohse P, Muller-Myhsok B, Limbersky O et al. Polymorphisms in the DLG5 and OCTN cation transporter genes in Crohn's disease. Gut 2005; 54(10):1421-1427.

(143) Tremelling M, Waller S, Bredin F, Greenfield S, Parkes M. Genetic variants in TNF-alpha but not DLG5 are associated with inflammatory bowel disease in a large United Kingdom cohort. Inflammatory Bowel Diseases 2006; 12(3):178-184.

(144) Pearce AV, Fisher SA, Prescott NJ, Onnie CM, Pattni R, Green P et al. Investigation of association of the DLG5 gene with phenotypes of inflammatory bowel disease in the British population. Int J Colorectal Dis 2007; 22(4):419-424.

(145) Browning BL, Huebner C, Petermann I, Demmers P, McCulloch A, Gearry RB et al. Association of DLG5 variants with inflammatory bowel disease in the New Zealand caucasian population and meta-analysis of the DLG5 R30Q variant. Inflammatory Bowel Diseases 2007; 13(9):1069-1076.

(146) Armuzzi A, Ahmad T, Ling KL, de Silva A, Cullen S, van Heel D et al. Genotype-phenotype analysis of the Crohn's disease susceptibility haplotype on chromosome 5q31. Gut 2003; 52(8):1133-1139.

(147) Giallourakis C, Stoll M, Miller K, Hampe J, Lander ES, Daly MJ et al. IBD5 is a general risk factor for inflammatory bowel disease: replication of association with Crohn disease and identification of a novel association with ulcerative colitis. American Journal of Human Genetics 2003; 73(1):205-211.

(148) Mirza MM, Fisher SA, King K, Cuthbert AP, Hampe J, Sanderson J et al. Genetic evidence for interaction of the 5q31 cytokine locus and the CARD15 gene in Crohn disease. American Journal of Human Genetics 2003; 72(4):1018-1022.

(149) Torkvist L, Noble CL, Lordal M, Sjoqvist U, Lindforss U, Nimmo ER et al. Contribution of the IBD5 locus to Crohn's disease in the Swedish population. Scandinavian Journal of Gastroenterology 2007; 42(2):200-206.

(150) Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z et al. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nature Genetics 2001; 29(2):223-228.

(151) Peltekova VD, Wintle RF, Rubin LA, Amos CI, Huang QQ, Gu XJ et al. Functional variants of OCTN cation transporter genes are associated with Crohn disease. Nature Genetics 2004; 36(5):471-475.

(152) Noble CL, Nimmo ER, Drummond H, Ho G-T, Tenesa A, Smith L et al. The contribution of OCTN1/2 variants within the IBD5 locus to disease susceptibility and severity in Crohn's disease. Gastroenterology 2005; 129(6):1854-1864.

(153) Babusukumar U, Wang T, McGuire E, Broeckel U, Kugathasan S. Contribution of OCTN variants within the IBD5 locus to pediatric onset Crohn's disease. American Journal of Gastroenterology 2006; 101(6):1354-1361.

(154) Silverberg MS, Duerr RH, Brant SR, Bromfield G, Datta LW, Jani N et al. Refined genomic localization and ethnic differences observed for the IBD5 association with Crohn's disease. Eur J Hum Genet 2007; 15(3):328-335.

(155) Vermeire S, Pierik M, Hlavaty T, Claessens G, van Schuerbeeck N, Joossens S et al. Association of organic cation transporter risk haplotype with perianal penetrating Crohn's disease but not with susceptibility to IBD. Gastroenterology 2005; 129(6):1845-1853.

(156) Halfvarson J. Genetics in twins with Crohn's disease: less pronounced than previously believed? Inflamm Bowel Dis 2011; 17(1):6-12.

(157) Mahid SS, Minor KS, Soto RE, Hornung CA, Galandiuk S. Smoking and inflammatory bowel disease: a meta-analysis. Mayo Clin Proc 2006; 81(11):1462-1471.

(158) Bridger S, Lee JC, Bjarnason I, Jones JE, MacPherson AJ. In siblings with similar genetic susceptibility for inflammatory bowel disease, smokers tend to develop Crohn's disease and non-smokers develop ulcerative colitis. Gut 2002; 51(1):21-25.

(159) Aldhous MC, Drummond HE, Anderson N, Smith LA, Arnott ID, Satsangi J. Does cigarette smoking influence the phenotype of Crohn's disease? Analysis using the Montreal classification. Am J Gastroenterol 2007; 102(3):577-588.

(160) Cosnes J, Beaugerie L, Carbonnel F, Gendre JP. Smoking cessation and the course of Crohn's disease: an intervention study. Gastroenterology 2001; 120(5):1093-1099.

(161) Thomas GA, Rhodes J, Mani V, Williams GT, Newcombe RG, Russell MA et al. Transdermal nicotine as maintenance therapy for ulcerative colitis. N Engl J Med 1995; 332(15):988-992.

(162) Pullan RD, Rhodes J, Ganesh S, Mani V, Morris JS, Williams GT et al. Transdermal nicotine for active ulcerative colitis. N Engl J Med 1994; 330(12):811-815.

(163) Ingram JR, Rhodes J, Evans BK, Thomas GA. Nicotine enemas for active Crohn's colitis: an open pilot study. Gastroenterol Res Pract 2008; 2008:237185.

(164) Ingram JR, Thomas GA, Rhodes J, Green JT, Hawkes ND, Swift JL et al. A randomized trial of nicotine enemas for active ulcerative colitis. Clin Gastroenterol Hepatol 2005; 3(11):1107-1114.

(165) Koutroubakis IE, Vlachonikolis IG. Appendectomy and the development of ulcerative colitis: results of a metaanalysis of published case-control studies. Am J Gastroenterol 2000; 95(1):171-176.

(166) Kaplan GG, Jackson T, Sands BE, Frisch M, Andersson RE, Korzenik J. The risk of developing Crohn's disease after an appendectomy: a meta-analysis. Am J Gastroenterol 2008; 103(11):2925-2931.

(167) O'Sullivan M. Symposium on 'The challenge of translating nutrition research into public health nutrition'. Session 3: Joint Nutrition Society and Irish Nutrition and Dietetic Institute Symposium on 'Nutrition and autoimmune disease'. Nutrition in Crohn's disease. Proc Nutr Soc 2009; 68(2):127-134.

(168) Sakamoto N, Kono S, Wakai K, Fukuda Y, Satomi M, Shimoyama T et al. Dietary risk factors for inflammatory bowel disease: a multicenter case-control study in Japan. Inflamm Bowel Dis 2005; 11(2):154-163.

(169) Amre DK, D'Souza S, Morgan K, Seidman G, Lambrette P, Grimard G et al. Imbalances in dietary consumption of fatty acids, vegetables, and fruits are associated with risk for Crohn's disease in children. Am J Gastroenterol 2007; 102(9):2016-2025.

(170) de Silva PS, Olsen A, Christensen J, Schmidt EB, Overvaad K, Tjonneland A et al. An association between dietary arachidonic acid, measured in adipose tissue, and ulcerative colitis. Gastroenterology 2010; 139(6):1912-1917.

(171) de Silva PS, Luben R, McTaggart A, Khaw KT, Hart AR. Dietary arachidonic acid and the aetiology of ulcerative colitis - data from a UK prospective cohort study, using 7-day food diaries. Digestive Disease Week , 632. 2011.
Ref Type: Abstract

(172) Strachan DP. Hay fever, hygiene, and household size. BMJ 1989; 299(6710):1259-1260.

(173) Armitage EL, Aldhous MC, Anderson N, Drummond HE, Riemersma RA, Ghosh S et al. Incidence of juvenile-onset Crohn's disease in Scotland: association with northern latitude and affluence. Gastroenterology 2004; 127(4):1051-1057.

(174) Declercq C, Gower-Rousseau C, Vernier-Massouille G, Salleron J, Balde M, Poirier G et al. Mapping of inflammatory bowel disease in northern France: spatial variations and relation to affluence. Inflamm Bowel Dis 2010; 16(5):807-812.

(175) Parsot C. Shigella spp. and enteroinvasive Escherichia coli pathogenicity factors. FEMS Microbiol Lett 2005; 252(1):11-18.

(176) Slack E, Hapfelmeier S, Stecher B, Velykoredko Y, Stoel M, Lawson MA et al. Innate and adaptive immunity cooperate flexibly to maintain host-microbiota mutualism. Science 2009; 325(5940):617-620.

(177) Sellon RK, Tonkonogy S, Schultz M, Dieleman LA, Grenther W, Balish E et al. Resident enteric bacteria are necessary for development of spontaneous colitis and immune system activation in interleukin-10-deficient mice. Infect Immun 1998; 66(11):5224-5231.

(178) Arnold GL, Beaves MR, Pryjdun VO, Mook WJ. Preliminary study of ciprofloxacin in active Crohn's disease. Inflamm Bowel Dis 2002; 8(1):10-15.

(179) D'Haens GR, Geboes K, Peeters M, Baert F, Penninckx F, Rutgeerts P. Early lesions of recurrent Crohn's disease caused by infusion of intestinal contents in excluded ileum. Gastroenterology 1998; 114(2):262-267.

(180) Porter CK, Tribble DR, Aliaga PA, Halvorson HA, Riddle MS. Infectious gastroenteritis and risk of developing inflammatory bowel disease. Gastroenterology 2008; 135(3):781-786.

(181) Gradel KO, Nielsen HL, Schonheyder HC, Ejlertsen T, Kristensen B, Nielsen H. Increased short- and long-term risk of inflammatory bowel disease after salmonella or campylobacter gastroenteritis. Gastroenterology 2009; 137(2):495-501.

(182) Leung AK, Robson WL, Davies HD. Traveler's diarrhea. Adv Ther 2006; 23(4):519-527.

(183) Darfeuille-Michaud A, Boudeau J, Bulois P, Neut C, Glasser AL, Barnich N et al. High prevalence of adherent-invasive Escherichia coli associated with ileal mucosa in Crohn's disease. Gastroenterology 2004; 127(2):412-421.

(184) Martin HM, Campbell BJ, Hart CA, Mpofu C, Nayar M, Singh R et al. Enhanced Escherichia coli adherence and invasion in Crohn's disease and colon cancer. Gastroenterology 2004; 127(1):80-93.

(185) Tabaqchali S, O'Donoghue DP, Bettelheim KA. Escherichia coli antibodies in patients with inflammatory bowel disease. Gut 1978; 19(2):108-113.

(186) Mow WS, Landers CJ, Steinhart AH, Feagan BG, Croitoru K, Seidman E et al. High-level serum antibodies to bacterial antigens are associated with antibiotic-induced clinical remission in Crohn's disease: a pilot study. Dig Dis Sci 2004; 49(7-8):1280-1286.

(187) Mow WS, Vasiliauskas EA, Lin YC, Fleshner PR, Papadakis KA, Taylor KD et al. Association of antibody responses to microbial antigens and complications of small bowel Crohn's disease. Gastroenterology 2004; 126(2):414-424.

(188) Arnott ID, Landers CJ, Nimmo EJ, Drummond HE, Smith BK, Targan SR et al. Sero-reactivity to microbial components in Crohn's disease is associated with disease severity and progression, but not NOD2/CARD15 genotype. Am J Gastroenterol 2004; 99(12):2376-2384.

(189) Rolhion N, Carvalho FA, rfeuille-Michaud A. OmpC and the sigma(E) regulatory pathway are involved in adhesion and invasion of the Crohn's disease-associated Escherichia coli strain LF82. Mol Microbiol 2007; 63(6):1684-1700.

(190) Glasser AL, Boudeau J, Barnich N, Perruchot MH, Colombel JF, rfeuille-Michaud A. Adherent invasive Escherichia coli strains from patients with Crohn's disease survive and replicate within macrophages without inducing host cell death. Infect Immun 2001; 69(9):5529-5537.

(191) Lapaquette P, Glasser AL, Huett A, Xavier RJ, rfeuille-Michaud A. Crohn's disease-associated adherent-invasive E. coli are selectively favoured by impaired autophagy to replicate intracellularly. Cell Microbiol 2010; 12(1):99-113.

(192) Stabel JR. Johne's disease: a hidden threat. J Dairy Sci 1998; 81(1):283-288.

(193) Chiodini RJ, Van Kruiningen HJ, Thayer WR, Merkal RS, Coutu JA. Possible role of mycobacteria in inflammatory bowel disease. I. An unclassified Mycobacterium species isolated from patients with Crohn's disease. Dig Dis Sci 1984; 29(12):1073-1079.

(194) Millar D, Ford J, Sanderson J, Withey S, Tizard M, Doran T et al. IS900 PCR to detect Mycobacterium paratuberculosis in retail supplies of whole pasteurized cows' milk in England and Wales. Appl Environ Microbiol 1996; 62(9):3446-3452.

(195) Mishina D, Katsel P, Brown ST, Gilberts EC, Greenstein RJ. On the etiology of Crohn disease. Proc Natl Acad Sci U S A 1996; 93(18):9816-9820.

(196) Sanderson JD, Moss MT, Tizard ML, Hermon-Taylor J. Mycobacterium paratuberculosis DNA in Crohn's disease tissue. Gut 1992; 33(7):890-896.

(197) Autschbach F, Eisold S, Hinz U, Zinser S, Linnebacher M, Giese T et al. High prevalence of Mycobacterium avium subspecies paratuberculosis IS900 DNA in gut tissues from individuals with Crohn's disease. Gut 2005; 54(7):944-949.

(198) Hulten K, El-Zimaity HM, Karttunen TJ, Almashhrawi A, Schwartz MR, Graham DY et al. Detection of Mycobacterium avium subspecies paratuberculosis in Crohn's diseased tissues by in situ hybridization. Am J Gastroenterol 2001; 96(5):1529-1535.

(199) Juste RA, Elguezabal N, Garrido JM, Pavon A, Geijo MV, Sevilla I et al. On the prevalence of M. avium subspecies paratuberculosis DNA in the blood of healthy individuals and patients with inflammatory bowel disease. PLoS One 2008; 3(7):e2537.

(200) Selby W, Pavli P, Crotty B, Florin T, Radford-Smith G, Gibson P et al. Two-year combination antibiotic therapy with clarithromycin, rifabutin, and clofazimine for Crohn's disease. Gastroenterology 2007; 132(7):2313-2319.

(201) Zhang FR, Huang W, Chen SM, Sun LD, Liu H, Li Y et al. Genomewide Association Study of Leprosy. New England Journal of Medicine 2009; 361(27):2609-2618.

(202) Sokol H, Pigneur B, Watterlot L, Lakhdari O, Bermudez-Humaran LG, Gratadoux JJ et al. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proc Natl Acad Sci U S A 2008; 105(43):16731-16736.

(203) Sokol H, Seksik P, Furet JP, Firmesse O, Nion-Larmurier I, Beaugerie L et al. Low counts of Faecalibacterium prausnitzii in colitis microbiota. Inflamm Bowel Dis 2009; 15(8):1183-1189.

(204) Berg RD. The indigenous gastrointestinal microflora. Trends Microbiol 1996; 4(11):430-435.

(205) Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L et al. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut 2006; 55(2):205-211.

(206) Kang S, Denman SE, Morrison M, Yu Z, Dore J, Leclerc M et al. Dysbiosis of fecal microbiota in Crohn's disease patients as revealed by a custom phylogenetic microarray. Inflamm Bowel Dis 2010; 16(12):2034-2042.

(207) Hooper LV. Do symbiotic bacteria subvert host immunity? Nat Rev Microbiol 2009; 7(5):367-374.

(208) Atuma C, Strugala V, Allen A, Holm L. The adherent gastrointestinal mucus gel layer: thickness and physical state in vivo. Am J Physiol Gastrointest Liver Physiol 2001; 280(5):G922-G929.

(209) Hilsden RJ, Meddings JB, Sutherland LR. Intestinal permeability changes in response to acetylsalicylic acid in relatives of patients with Crohn's disease. Gastroenterology 1996; 110(5):1395-1403.

(210) Buisine MP, Desreumaux P, Leteurtre E, Copin MC, Colombel JF, Porchet N et al. Mucin gene expression in intestinal epithelial cells in Crohn's disease. Gut 2001; 49(4):544-551.

(211) Kerr SM, Liewald DC, Campbell A, Taylor K, Wild SH, Newby D et al. Generation Scotland: Donor DNA Databank; A control DNA resource. BMC Med Genet 2010; 11:166.

(212) The International HapMap Consortium. A haplotype map of the human genome. Nature 2005; 437(7063):1299-1320.

(213) Clark AG. Genomics of the evolutionary process. Trends Ecol Evol 2006; 21(6):316-321.

(214) Abdi H. Bonferroni Test. In: Salkind NJ, editor. *Encyclopedia of Measurement and Statistics*. Sage; 2007. 103-107.

(215) de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. Nature Genetics 2005; 37(11):1217-1223.

(216) Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol 2002; 155(5):478-484.

(217) Kaplan E, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 1958; 53:457-481.

(218) Oberhuber G, Stangl PC, Vogelsang H, Schober E, Herbst F, Gasche C. Significant association of strictures and internal fistula formation in Crohn's disease. Virchows Arch 2000; 437(3):293-297.

(219) Kelly JK, Preshaw RM. Origin of fistulas in Crohn's disease. J Clin Gastroenterol 1989; 11(2):193-196.

(220) Kahn E, Markowitz J, Blomquist K, Daum F. The morphologic relationship of sinus and fistula formation to intestinal stenoses in children with Crohn's disease. Am J Gastroenterol 1993; 88(9):1395-1398.

(221) Russel MG, Volovics A, Schoon EJ, van Wijlick EH, Logan RF, Shivananda S et al. Inflammatory bowel disease: is there any relation between smoking status and disease presentation? European Collaborative IBD Study Group. Inflamm Bowel Dis 1998; 4(3):182-186.

(222) Picco MF, Bayless TM. Tobacco consumption and disease duration are associated with fistulizing and stricturing behaviors in the first 8 years of Crohn's disease. Am J Gastroenterol 2003; 98(2):363-368.

(223) Cosnes J, Cattan S, Blain A, Beaugerie L, Carbonnel F, Parc R et al. Long-Term Evolution of Disease Behavior of Crohn's Disease. Inflammatory Bowel Diseases 2002; 8(4):244-250.

(224) Tarrant KM, Barclay ML, Frampton CM, Gearry RB. Perianal disease predicts changes in Crohn's disease phenotype-results of a population-based study of inflammatory bowel disease phenotype. Am J Gastroenterol 2008; 103(12):3082-3093.

(225) Louis E, Michel V, Hugot JP, Reenaers C, Fontaine F, Delforge M et al. Early development of stricturing or penetrating pattern in Crohn's disease is influenced by disease location, number of flares, and smoking but not by NOD2/CARD15 genotype. Gut 2003; 52(4):552-557.

(226) Ramadas AV, GUNESH S, Thomas GA, Williams GT, HAWTHORNE AB. Natural history of Crohn's disease in a population-based cohort from Cardiff (1986-2003): a study of changes in medical treatment and surgical resection rates. Gut 2010; 59(9):1200-1206.

(227) Sands BE, Arsenault JE, Rosen MJ, Alsahli M, Bailen L, Banks P et al. Risk of early surgery for Crohn's disease: implications for early treatment strategies. Am J Gastroenterol 2003; 98(12):2712-2718.

(228) Solberg IC, Vatn MH, Hoie O, Stray N, Sauar J, Jahnsen J et al. Clinical course in Crohn's disease: results of a Norwegian population-based ten-year follow-up study. Clin Gastroenterol Hepatol 2007; 5(12):1430-1438.

(229) Binion DG, Theriot KR, Shidham S, Lundeen S, Hatoum O, Lim HJ et al. Clinical factors contributing to rapid reoperation for Crohn's disease patients undergoing resection and/or strictureplasty. J Gastrointest Surg 2007; 11(12):1692-1698.

(230) Lautenbach E, Berlin JA, Lichtenstein GR. Risk factors for early postoperative recurrence of Crohn's disease. Gastroenterology 1998; 115(2):259-267.

(231) D'Haens GR. Top-down therapy for IBD: rationale and requisite evidence. Nat Rev Gastroenterol Hepatol 2010; 7(2):86-92.

(232) Lees CW, Ali AI, Thompson AI, Ho GT, Forsythe RO, Marquez L et al. The safety profile of anti-tumour necrosis factor therapy in inflammatory bowel disease in clinical practice: analysis of 620 patient-years follow-up. Aliment Pharmacol Ther 2009; 29(3):286-297.

(233) Ho GT, Mowat A, Potts L, Cahill A, Mowat C, Lees CW et al. Efficacy and complications of adalimumab treatment for medically-refractory Crohn's disease: analysis of nationwide experience in Scotland (2004-2008). Aliment Pharmacol Ther 2009; 29(5):527-534.

(234) Beaugerie L, Seksik P, Nion-Larmurier I, Gendre JP, Cosnes J. Predictors of Crohn's disease. Gastroenterology 2006; 130(3):650-656.

(235) Loly C, Belaiche J, Louis E. Predictors of severe Crohn's disease. Scand J Gastroenterol 2008; 43(8):948-954.

(236) Pariente B, Cosnes J, Danese S, Sandborn WJ, Lewin M, Fletcher JG et al. Development of the Crohn's disease digestive damage score, the Lemann score. Inflamm Bowel Dis 2011; 17(6):1415-1422.

(237) Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 2009; 460(7256):748-752.

(238) Gianfrancesco F, Esposito T, Ombra MN, Forabosco P, Maninchedda G, Fattorini M et al. Identification of a novel gene and a common variant associated with uric acid nephrolithiasis in a Sardinian genetic isolate. Am J Hum Genet 2003; 72(6):1479-1491.

(239) Felder JB, Korelitz BI, Rajapakse R, Schwarz S, Horatagis AP, Gleim G. Effects of nonsteroidal antiinflammatory drugs on inflammatory bowel disease: a case-control study. Am J Gastroenterol 2000; 95(8):1949-1954.

(240) Kabashima K, Saji T, Murata T, Nagamachi M, Matsuoka T, Segi E et al. The prostaglandin receptor EP4 suppresses colitis, mucosal damage and CD4 cell activation in the gut. J Clin Invest 2002; 109(7):883-893.

(241) Breslow DK, Collins SR, Bodenmiller B, Aebersold R, Simons K, Shevchenko A et al. Orm family proteins mediate sphingolipid homeostasis. Nature 2010; 463(7284):1048-1053.

(242) Lingwood D, Simons K. Lipid rafts as a membrane-organizing principle. Science 2010; 327(5961):46-50.

(243) Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 2007; 448(7152):470-473.

(244) Weersma RK, Stokkers PC, van Bodegraven AA, van Hogezand RA, Verspaget HW, De Jong DJ et al. Molecular prediction of disease risk and severity in a large Dutch Crohn's disease cohort. Gut 2009; 58(3):388-395.

(245) Clausen H, Bennett EP. A family of UDP-GalNAc: Polypeptide N-acetylgalactosaminyl-transferases control the initiation of mucin-type O-linked glycosylation. Glycobiology 1996; 6(6):635-646.

(246) Ten Hagen KG, Fritz TA, Tabak LA. All in the family: the UDP-GalNAc : polypeptide N-acetylgalactosaminyltransferases. Glycobiology 2003; 13(1):1R-16R.

(247) Guda K, Moinova H, He J, Jamison O, Ravi L, Natale L et al. Inactivating germ-line and somatic mutations in polypeptide N-acetylgalactosaminyltransferase 12 in human colon cancers. Proc Natl Acad Sci U S A 2009; 106(31):12921-12925.

(248) Ishikawa M, Kitayama J, Nariko H, Kohno K, Nagawa H. The expression pattern of UDP-N-acetyl-alpha-d-galactosamine: polypeptide N-acetylgalactosaminyl transferase-3 in early gastric carcinoma. J Surg Oncol 2004; 86(1):28-33.

(249) Ishikawa M, Kitayama J, Kohno K, Nagawa H. The expression pattern of UDP-N-acetyl-alpha-D-galactosamine-polypeptide N-acetyl-galactosaminyl transferase-3 in squamous cell carcinoma of the esophagus. Pathobiology 2005; 72(3):139-145.

(250) Yamamoto S, Nakamori S, Tsujie M, Takahashi Y, Nagano H, Dono K et al. Expression of uridine diphosphate N-acetyl-alpha-D-galactosamine: polypeptide N-acetylgalactosaminyl transferase 3 in adenocarcinoma of the pancreas. Pathobiology 2004; 71(1):12-18.

(251) Shibao K, Izumi H, Nakayama Y, Ohta R, Nagata N, Nomoto M et al. Expression of UDP-N-acetyl-alpha-D-galactosamine-polypeptide galNAc

N-acetylgalactosaminyl transferase-3 in relation to differentiation and prognosis in patients with colorectal carcinoma. Cancer 2002; 94(7):1939-1946.

(252) Chefetz I, Sprecher E. Familial tumoral calcinosis and the role of O-glycosylation in the maintenance of phosphate homeostasis. Biochim Biophys Acta 2009; 1792(9):847-852.

(253) Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 2008; 40(2):161-169.

(254) Schjoldager KT, Vester-Christensen MB, Bennett EP, Levery SB, Schwientek T, Yin W et al. O-glycosylation modulates proprotein convertase activation of angiopoietin-like protein 3: possible role of polypeptide GalNAc-transferase-2 in regulation of concentrations of plasma lipids. J Biol Chem 2010; 285(47):36293-36303.

(255) Fields S, Song O. A novel genetic system to detect protein-protein interactions. Nature 1989; 340(6230):245-246.

(256) Inoue K, Mallakin A, Frazier DP. Dmp1 and tumor suppression. Oncogene 2007; 26(30):4329-4335.

(257) Lake JA, Carr J, Feng F, Mundy L, Burrell C, Li P. The role of Vif during HIV-1 infection: interaction with novel host cellular factors. Journal of Clinical Virology 2003; 26(2):143-152.

(258) Hattori T, Baba K, Matsuzaki S, Honda A, Miyoshi K, Inoue K et al. A novel DISC1-interacting partner DISC1-binding zinc-finger protein: implication in the modulation of DISC1-dependent neurite outgrowth. Molecular Psychiatry 2007; 12(4):398-407.

(259) Yamamoto-Furusho JK, Barnich N, Xavier R, Hisamatsu T, Podolsky DK. Centaurin beta1 down-regulates nucleotide-binding oligomerization domains 1- and 2-dependent NF-kappaB activation. J Biol Chem 2006; 281(47):36060-36070.

(260) Barnich N, Hisamatsu T, Aguirre JE, Xavier R, Reinecker HC, Podolsky DK. GRIM-19 interacts with nucleotide oligomerization domain 2 and serves as downstream effector of anti-bacterial function in intestinal epithelial cells. J Biol Chem 2005; 280(19):19021-19026.

(261) Nimmo ER, Phillips AM, Stevens C, Smith A, Drummond HE, Noble CL et al. Analysis of protein-protein and gene-gene interactions implicates

TLE1 as a critical modifier of NOD2 effect in Crohn's disease. Gastroenterology. In press 2011.

(262) Bennett EP, Weghuis DO, Merkx G, van Kessel AG, Eiberg H, Clausen H. Genomic organization and chromosomal localization of three members of the UDP-N-acetylgalactosamine: polypeptide N-acetylgalactosaminyltransferase family. Glycobiology 1998; 8(6):547-555.

(263) White T, Bennett EP, Takio K, Sorensen T, Bonding N, Clausen H. Purification and Cdna Cloning of A Human Udp-N-Acetyl-Alpha-D-Galactosamine-Polypeptide N-Acetylgalactosaminyltransferase. J Biol Chem 1995; 270(41):24156-24165.

(264) Ogura Y, Lala S, Xin W, Smith E, Dowds TA, Chen FF et al. Expression of NOD2 in Paneth cells: A possible link to Crohn's ileitis. Gut 2003; 52(11):1591-1597.

(265) Dinter A, Berger EG. Golgi-disturbing agents. Histochem Cell Biol 1998; 109(5-6):571-590.

(266) Tartakoff AM. Perturbation of vesicular traffic with the carboxylic ionophore monensin. Cell 1983; 32(4):1026-1028.

(267) Roth J, Wang Y, Eckhardt AE, Hill RL. Subcellular localization of the UDP-N-acetyl-D-galactosamine: polypeptide N-acetylgalactosaminyltransferase-mediated O-glycosylation reaction in the submaxillary gland. Proc Natl Acad Sci U S A 1994; 91(19):8935-8939.

(268) Zhang GF, Driouich A, Staehelin LA. Effect of monensin on plant Golgi: re-examination of the monensin-induced changes in cisternal architecture and functional activities of the Golgi apparatus of sycamore suspension-cultured cells. J Cell Sci 1993; 104 (Pt 3):819-831.

(269) McCool DJ, Forstner JF, Forstner GG. Regulated and unregulated pathways for MUC2 mucin secretion in human colonic LS180 adenocarcinoma cells are distinct. Biochem J 1995; 312 (Pt 1):125-133.

(270) Kanno H, Horikawa Y, Hodges RR, Zoukhri D, Shatos MA, Rios JD et al. Cholinergic agonists transactivate EGFR and stimulate MAPK to induce goblet cell secretion. Am J Physiol Cell Physiol 2003; 284(4):C988-C998.

(271) Rosenstiel P, Fantini M, Brautigam K, Kuhbacher T, Waetzig GH, Seegert D et al. TNF-alpha and IFN-gamma regulate the expression of the NOD2 (CARD15) gene in human intestinal epithelial cells. Gastroenterology 2003; 124(4):1001-1009.

(272) Hisamatsu T, Suzuki M, Reinecker HC, Nadeau WJ, McCormick BA, Podolsky DK. CARD15/NOD2 functions as an antibacterial factor in human intestinal epithelial cells. Gastroenterology 2003; 124(4):993-1000.

(273) Takahashi Y, Isuzugawa K, Murase Y, Imai M, Yamamoto S, Iizuka M et al. Up-regulation of NOD1 and NOD2 through TLR4 and TNF-alpha in LPS-treated murine macrophages. J Vet Med Sci 2006; 68(5):471-478.

(274) Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 2001; 25(4):402-408.

(275) Pigny P, Guyonnet-Duperat V, Hill AS, Pratt WS, Galiegue-Zouitina S, d'Hooge MC et al. Human mucin genes assigned to 11p15.5: identification and organization of a cluster of genes. Genomics 1996; 38(3):340-352.

(276) Fox MF, Lahbib F, Pratt W, Attwood J, Gum J, Kim Y et al. Regional Localization of the Intestinal Mucin Gene Muc3 to Chromosome 7Q22. Annals of Human Genetics 1992; 56:281-287.

(277) Gross MS, GuyonnetDuperat V, Porchet N, Bernheim A, Aubert JP, Nguyen VC. Mucin-4 (Muc4) Gene - Regional Assignment (3Q29) and Rflp Analysis. Annales de Genetique 1992; 35(1):21-26.

(278) Swallow DM, Gendler S, Griffiths B, Kearney A, Povey S, Sheer D et al. The Hypervariable Gene Locus Pum, Which Codes for the Tumor Associated Epithelial Mucins, Is Located on Chromosome-1, Within the Region 1Q21-24. Annals of Human Genetics 1987; 51:289-294.

(279) Hang HC, Bertozzi CR. The chemistry and biology of mucin-type O-linked glycosylation. Bioorg Med Chem 2005; 13(17):5021-5034.

(280) Moehle C, Ackermann N, Langmann T, Aslanidis C, Kel A, Kel-Margoulis O et al. Aberrant intestinal expression and allelic variants of mucin genes associated with inflammatory bowel disease. Journal of Molecular Medicine 2006; 84(12):1055-1066.

(281) Deplancke B, Gaskins HR. Microbial modulation of innate defense: goblet cells and the intestinal mucus layer. Am J Clin Nutr 2001; 73(6):1131S-1141S.

(282) Allen A, Hutton DA, Pearson JP. The MUC2 gene product: a human intestinal mucin. International Journal of Biochemistry & Cell Biology 1998; 30(7):797-801.

(283) Byrd JC, Bresalier RS. Mucins and mucin binding proteins in colorectal cancer. Cancer Metastasis Rev 2004; 23(1-2):77-99.

(284) Williams SJ, Munster DJ, Quin RJ, Gotley DC, McGuckin MA. The MUC3 gene encodes a transmembrane mucin and is alternatively spliced. Biochemical and Biophysical Research Communications 1999; 261(1):83-89.

(285) Pratt WS, Crawley S, Hicks J, Ho J, Nash M, Kim YS et al. Multiple transcripts of MUC3: Evidence for two genes, MUC3A and MUC3B. Biochemical and Biophysical Research Communications 2000; 275(3):916-923.

(286) Kyo K, Parkes M, Takei Y, Nishimori H, Vyas P, Satsangi J et al. Association of ulcerative colitis with rare VNTR alleles of the human intestinal mucin gene, MUC3. Hum Mol Genet 1999; 8(2):307-311.

(287) Williams SJ, McGuckin MA, Gotley DC, Eyre HJ, Sutherland GR, Antalis TM. Two novel mucin genes down-regulated in colorectal cancer identified by differential display. Cancer Res 1999; 59(16):4083-4089.

(288) Williams SJ, Wreschner DH, Tran M, Eyre HJ, Sutherland GR, McGuckin MA. MUC13, a novel human cell surface mucin expressed by epithelial and hemopoietic cells. J Biol Chem 2001; 276(21):18327-18336.

(289) Gum JR, Crawley SC, Hicks JW, Szymkowski DE, Kim YS. MUC17, a novel membrane-tethered mucin. Biochemical and Biophysical Research Communications 2002; 291(3):466-475.

(290) Chen Y, Zhao YH, Kalaslavadi TB, Halmati E, Nehrke K, Le AD et al. Genome-wide search and identification of a novel gel-forming mucin MUC19/Muc19 in glandular tissues. American Journal of Respiratory Cell and Molecular Biology 2004; 30(2):155-165.

(291) Kim YS, Ho SB. Intestinal goblet cells and mucins in health and disease: recent insights and progress. Curr Gastroenterol Rep 2010; 12(5):319-330.

(292) Ogata S, Uehara H, Chen A, Itzkowitz SH. Mucin gene expression in colonic tissues and cell lines. Cancer Res 1992; 52(21):5971-5978.

(293) Weiss AA, Babyatsky MW, Ogata S, Chen A, Itzkowitz SH. Expression of MUC2 and MUC3 mRNA in human normal, malignant, and inflammatory intestinal tissues. J Histochem Cytochem 1996; 44(10):1161-1166.

(294) Velcich A, Yang W, Heyer J, Fragale A, Nicholas C, Viani S et al. Colorectal cancer in mice genetically deficient in the mucin Muc2. Science 2002; 295(5560):1726-1729.

(295) Yonezawa S, Sato E. Expression of mucin antigens in human cancers and its relationship with malignancy potential. Pathol Int 1997; 47(12):813-830.

(296) Hoebler C, Gaudier E, de Coppet P, Rival M, Cherbut C. MUC genes are differently expressed during onset and maintenance of inflammation in dextran sodium sulfate-treated mice. Digestive Diseases & Sciences 2006; 51(2):381-389.

(297) Faure M, Moennoz D, Montigon F, Mettraux C, Mercier S, Schiffrin EJ et al. Mucin production and composition is altered in dextran sulfate sodium-induced colitis in rats. Dig Dis Sci 2003; 48(7):1366-1373.

(298) Van der SM, De Koning BA, De Bruijn AC, Velcich A, Meijerink JP, Van Goudoever JB et al. Muc2-deficient mice spontaneously develop colitis, indicating that MUC2 is critical for colonic protection. Gastroenterology 2006; 131(1):117-129.

(299) Pullan RD, Thomas GAO, Rhodes M, Newcombe RG, Williams GT, Allen A et al. Thickness of Adherent Mucus Gel on Colonic Mucosa in Humans and Its Relevance to Colitis. Gut 1994; 35(3):353-359.

(300) McCormick DA, Horton LW, Mee AS. Mucin depletion in inflammatory bowel disease. J Clin Pathol 1990; 43(2):143-146.

(301) Theodossi A, Spiegelhalter DJ, Jass J, Firth J, Dixon M, Leader M et al. Observer variation and discriminatory value of biopsy features in inflammatory bowel disease. Gut 1994; 35(7):961-968.

(302) Van Klinken BJ, Van der Wal JW, Einerhand AW, Buller HA, Dekker J. Sulphation and secretion of the predominant secretory human colonic mucin MUC2 in ulcerative colitis. Gut 1999; 44(3):387-393.

(303) Kyo K, Muto T, Nagawa H, Lathrop GM, Nakamura Y. Associations of distinct variants of the intestinal mucin gene MUC3A with ulcerative colitis and Crohn's disease. Journal of Human Genetics 2001; 46(1):5-20.

(304) Satsangi J, Parkes M, Louis E, Hashimoto L, Kato N, Welsh K et al. Two stage genome-wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosomes 3, 7 and 12. Nature Genetics 1996; 14(2):199-202.

(305) Vincent A, Perrais M, Desseyn JL, Aubert JP, Pigny P, Van S, I. Epigenetic regulation (DNA methylation, histone modifications) of the 11p15 mucin genes (MUC2, MUC5AC, MUC5B, MUC6) in epithelial cancer cells. Oncogene 2007; 26(45):6566-6576.

(306) Kitamoto S, Yamada N, Yokoyama S, Houjou I, Higashi M, Yonezawa S. Promoter hypomethylation contributes to the expression of MUC3A in cancer cells. Biochem Biophys Res Commun 2010; 397(2):333-339.

(307) Burger-van PN, Vincent A, Puiman PJ, Van der SM, Bouma J, Boehm G et al. The regulation of intestinal mucin MUC2 expression by short-chain fatty acids: implications for epithelial protection. Biochem J 2009; 420(2):211-219.

(308) Li JD, Feng W, Gallup M, Kim JH, Gum J, Kim Y et al. Activation of NF-kappaB via a Src-dependent Ras-MAPK-pp90rsk pathway is required for Pseudomonas aeruginosa-induced mucin overproduction in epithelial cells. Proc Natl Acad Sci U S A 1998; 95(10):5718-5723.

(309) Ahn DH, Crawley SC, Hokari R, Kato S, Yang SC, Li JD et al. TNF-alpha activates MUC2 transcription via NF-kappaB but inhibits via JNK activation. Cell Physiol Biochem 2005; 15(1-4):29-40.

(310) Radhakrishnan P, Halagowder D, Devaraj SN. Altered expression of MUC2 and MUC5AC in response to Shigella infection, an in vivo study. Biochim Biophys Acta 2007; 1770(6):884-889.

(311) Corfield AP, Myerscough N, Bradfield N, Corfield CA, Gough M, Clamp JR et al. Colonic mucins in ulcerative colitis: evidence for loss of sulfation. Glycoconj J 1996; 13(5):809-822.

(312) Png CW, Linden SK, Gilshenan KS, Zoetendal EG, McSweeney CS, Sly LI et al. Mucolytic Bacteria With Increased Prevalence in IBD Mucosa Augment In Vitro Utilization of Mucin by Other Bacteria. Am J Gastroenterol 2010.

(313) Hollingsworth MA, Swanson BJ. Mucins in cancer: Protection and control of the cell surface. Nature Reviews Cancer 2004; 4(1):45-60.

(314) Rousseau K, Kirkham S, Johnson L, Fitzpatrick B, Howard M, Adams EJ et al. Proteomic analysis of polymeric salivary mucins: no evidence for MUC19 in human saliva. Biochemical Journal 2008; 413:545-552.

(315) Yu DF, Chen Y, Han JM, Zhang H, Chen XP, Zou WJ et al. MUC19 expression in human ocular surface and lacrimal gland and its alteration in

Sjogren syndrome patients. Experimental Eye Research 2008; 86(2):403-411.

(316) Kerschner JE. Mucin gene expression in human middle ear epithelium. Laryngoscope 2007; 117(9):1666-1676.

(317) Paisan-Ruiz C, Jain S, Evans EW, Gilks WP, Simon J, van der BM et al. Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. Neuron 2004; 44(4):595-600.

(318) Zimprich A, Biskup S, Leitner P, Lichtner P, Farrer M, Lincoln S et al. Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. Neuron 2004; 44(4):601-607.

(319) Smith WW, Pei Z, Jiang H, Moore DJ, Liang Y, West AB et al. Leucine-rich repeat kinase 2 (LRRK2) interacts with parkin, and mutant LRRK2 induces neuronal degeneration. Proc Natl Acad Sci U S A 2005; 102(51):18676-18681.

(320) Plowey ED, Cherra SJ, III, Liu YJ, Chu CT. Role of autophagy in G2019S-LRRK2-associated neurite shortening in differentiated SH-SY5Y cells. J Neurochem 2008; 105(3):1048-1056.

(321) Alegre-Abarrategui J, Wade-Martins R. Parkinson disease, LRRK2 and the endocytic-autophagic pathway. Autophagy 2009; 5(8):1208-1210.

(322) Gardet A, Benita Y, Li C, Sands BE, Ballester I, Stevens C et al. LRRK2 is involved in the IFN-gamma response and host response to pathogens. J Immunol 2010; 185(9):5577-5585.

(323) Cadwell K, Liu JY, Brown SL, Miyoshi H, Loh J, Lennerz JK et al. A key role for autophagy and the autophagy gene Atg16l1 in mouse and human intestinal Paneth cells. Nature 2008; 456(7219):259-263.

(324) Vermeire S, Van AG, Rutgeerts P. Review article: Altering the natural history of Crohn's disease--evidence for and against current therapies. Aliment Pharmacol Ther 2007; 25(1):3-12.

(325) Dubinsky MC, Lin YC, Dutridge D, Picornell Y, Landers CJ, Farrior S et al. Serum immune responses predict rapid disease progression among children with Crohn's disease: immune responses predict disease progression. Am J Gastroenterol 2006; 101(2):360-367.